UNIVERSIDADE DE BRASÍLIA - UNB

INSTITUTO DE LETRAS – IL

DEPARTAMENTO DE LÍNGUAS ESTRANGEIRAS E TRADUÇÃO - LET

CURSO DE LÍNGUAS ESTRANGEIRAS APLICADAS – LEA - MSI

GLOSSÁRIO MULTILÍNGUE ONLINE SOBRE MIGRAÇÃO E REFÚGIO: UMA PROPOSTA PARA TRADUTORES E INTÉRPRETES

ANNA BEATRIZ DIMAS FURTADO

Brasília - DF

2019

ANNA BEATRIZ DIMAS FURTADO

GLOSSÁRIO MULTILÍNGUE ONLINE SOBRE MIGRAÇÃO E REFÚGIO: UMA PROPOSTA PARA TRADUTORES E INTÉRPRETES

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação (LEA-MSI), sob orientação da Profa. Dra. Elisa Duarte Teixeira, da Universidade de Brasília (UnB).

BRASÍLIA, DF

2019

2019

ANNA BEATRIZ DIMAS FURTADO

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de bacharel em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação (LEA-MSI), sob orientação da Profa. Dra. Elisa Duarte Teixeira, da Universidade de Brasília (UnB).

Aprovado em ____/ ____/ _____.

_____

Profa. Dra. Elisa Duarte Teixeira

Universidade de Brasília

Orientadora

_____

Prof. Dr Cláudio Corrêa e Castro Gonçalves

Universidade de Brasília

Avaliador

_____

Prof. Dr Thiago Blanch Pires

Universidade de Brasília

Avaliador

BRASÍLIA, DF

2019

À Graziela e Rubens, meus pais. Aos meus familiares, amigos e professores. Mas, principalmente, aos intérpretes e refugiados - que esse trabalho possa auxilia-los de alguma maneira.

AGRADECIMENTOS

Inicialmente gostaria de dizer que este trabalho só foi possível porque tive ajuda de muitas pessoas, e quando digo muitas, me refiro a muitas mesmo! Então, vamos lá...

Primeiramente, gostaria de agradecer a Deus por ter me feito teimosa e persistente, qualidade cuja ausência afetaria complemente este trabalho. Em segundo lugar, meus pais, Rubens e Graziela, por sempre me incentivarem a estudar e por me apoiarem mesmo quando vocês não entendiam o porquê de eu estar aprendendo a quarta ou quinta língua. Obrigada por terem fé em mim mesmo quando eu não tinha. Aos meus irmãos, Juninho e Gabi por fazerem parte da minha vida, mesmo quando eu precisava de silêncio para fazer este, ou (um milhão de) outros trabalhos. À vovó Sueli, e às tias Regina e Rejane, que sempre me incentivaram e cuidaram para que eu tivesse a melhor educação possível.

Agradeço ao Rafa, pela paciência e o carinho, especialmente nos momentos em que eu já acreditava que nada daria mais certo. Obrigada pelas revisões, pelas xícaras de vitamina de banana ou simplesmente por aceitar a ir para a biblioteca e ficar até às três da matina me fazendo companhia. À Adriana e Mauro, por me emprestarem uma casa de descanso quando eu ficava muito exausta, mas especialmente à Adriana, por me emprestar um laptop quando o meu me abandonou; sem sua generosidade, este trabalho não teria sido completo.

Obrigada à Universidade de Brasília e sua política de bolsas de socio-vulnerabilidade, sem a qual eu não teria conseguido concluir este curso.

Abraço especial para a professora Susana Martínez, por ter identificado potencial em mim ainda caloura, me convidado a participar do grupo de pesquisa MOBILANG e me apresentado à professora Sabine Gorovitz (uma mulher incrível!), quem me iniciou no mundo da iniciação científica, inspirou todo este trabalho e me apresentou a minha orientadora.

Falando em orientadora, não posso deixar de agradecê-la! Elisa, muito obrigada pela confiança (perdão pelos atrasos, foi por uma boa causa!), por me apresentar a palavra da Linguística de Corpus (da qual espero não me desviar por enquanto), por me acolher nas correrias da vida, me encorajar a seguir na pesquisa e por me inspirar. Quando eu crescer, eu quero ser igual a você!

Ao prof. Cláudio Corrêa, coorientador deste projeto, sem o qual este trabalho jamais teria sido feito. Obrigada por plantar a sementinha do Python na disciplina de Língua e Programação. Agradeço pelas tardes de programação nas quais apanhamos do

Google, mas onde nos divertimos e rimos muito. Levarei seus conselhos e orientações para a vida inteira.

A todos os professores do LEA-MSI que me permitiram voar bem alto durante essa jornada. Em especial aos professores Marcos Carneiro e Thiago Blanch, que me permitiram ser monitora em suas disciplinas, o que me deu a oportunidade de me aprofundar com mais confiança nos meus projetos e afetou diretamente a qualidade deste trabalho.

À minha amiga Mônica, que conheci na disciplina de mestrado ministrada pela Elisa, por fazer parte dessa jornada que é fazer glossários direcionados por LC. Obrigada pelos dias de diversão, os de discussão dos trabalhos e pela nossa aventura em Maceió.

Aos meus amigos e colegas de curso do LEA-MSI, Jeferson, Augusto, Janaína, Anahy, Gustavo, Renan e Bianca, que embarcaram comigo nesse curso e me aguentaram nos momentos de estresse (que foram muitos!), nos momentos de diversão e amizade, que ficarão para todo sempre guardados no meu coração.

Aos meus colegas e professores dos grupos de pesquisas e extensão MOBILANG e TermTraDiCo, em especial a Letícia, Gustavo e Fernanda, por compartilharem as tarefas e fazerem parte da implementação do Banco de Intérpretes do MOBILANG. À professora Carolina Calvo Capilla, sempre pronta para ajudar sempre que enfrentássemos alguma dificuldade no desempenho das tarefas (ou na vida!).

Aos meus amigos da geofísica e de fora da UnB, que compreenderam (ou não), mas aceitaram as ausências muitas vezes prolongadas, porque eu estava simplesmente devastada!

À empresa júnior Quimera, por criar o ambiente necessário ao meu desenvolvimento profissional sem o qual eu não teria a experiência que tenho hoje, e que me deu a possibilidade de conhecer a Vic, May, Giovanne, Ingrid e Mari (e o Rafa!) e de também fazer reuniões infinitas.

Finalmente (e ufa!), espero não ter esquecido ninguém, mas agradeço de coração a todos que contribuíram de alguma forma com esse trabalho direta ou indiretamente, seja com pequenos comentários ou me ouvindo desabafar sobre as dificuldades. Esse trabalho tem um pedacinho de cada um de vocês!

“No struggle can ever succeed without women participating side by side with men. There are two powers in the world; one is the sword and the other is the pen. There is a third power stronger than both, that of women.”

Malala Yousafzai

RESUMO

Todos os anos, milhares de pessoas migram de seus locais de residência em busca de melhores oportunidades de vida. Em 2018, 68,5 milhões de pessoas foram forçadas a deixar seus lares; desses, 25,4 milhões são refugiados e 3,1 são solicitantes de refúgio. No Brasil, batemos o recorde de solicitações em 2017: 33.866. É nesse contexto que se insere o MOBILANG (Mobilidades e Línguas em Contato), grupo de pesquisa que visa investigar os contatos linguísticos ocasionados pela mobilidade humana e prover soluções linguísticas para os imigrantes e refugiados no Brasil. Com o intuito de auxiliar intérpretes voluntários e tradutores trabalhando nesse âmbito, uma das soluções previstas pelo Grupo foi a criação de um material de referência online sobre o tema. Nesse sentido, o objetivo do presente trabalho foi estabelecer as bases necessárias para disponibilizar o Glossário Multilíngue Online sobre Migração e Refúgio, sistematizado a partir do **Co**rpus **M**ultilíngue de **Mi**gração e **Re**fúgio (COMMIRE), que compilamos ao longo de quase três anos de Iniciação Científica. Desenvolvido em parceria com o grupo de pesquisa TermiTraDiCo (Terminologia e Tradução Direcionadas por Corpus), o Glossário se apoiou nos pressupostos teóricos da Linguística de Corpus para extrair Unidades de Tradução Especializadas (UTEs) do referido corpus utilizando o programa Sketch Engine. As Unidades mais recorrentes foram, então, organizadas em um banco de dados com fichas YAML e, por meio de um sistema de busca em Python, são disponibilizadas online para consulta. Esperamos, com esse primeiro passo, ter contribuído para a criação de um material de apoio ágil e confiável, que poderá crescer conforme a demanda e a disponibilidade de contribuições, para amparar linguisticamente tanto intérpretes e tradutores da área quanto os próprios migrantes.

Palavras-chave: Glossário multilíngue. Terminologia direcionada por corpus. Banco de dados terminológico. Linguística de Corpus. Imigração e Refúgio.

ABSTRACT

Every year, thousands of people migrate from where they live in search of better life opportunities. In 2018, 68.5 million people were forced to leave their homes; from these, 25.4 million are refugees and 3.1 are asylum seekers. In Brazil, we have beaten the record of asylum claims in 2017: we received 33.866. In this context, MOBILANG (Mobilities and Languages in Contact) research group aims at investigating linguistic contacts emerging from human mobility and providing linguistic solutions for immigrants and refugees in Brazil. In order to assist volunteer interpreters and translators working in this field, one of the solutions proposed by the Group was the creation of an online reference material on the topic. In this sense, the goal of the present work is to establish the necessary foundations to make the Online Multilingual Glossary on Migration and Asylum available. This glossary was systematized from the Multilingual Corpus on Migration and Asylum (COMMIRE, in Portuguese), which we compiled throughout three years of undergraduate research mentorship. Developed in partnership with the research group TermiTraDiCo (Corpus-driven Terminology and Translation), the Glossary was based on the theoretical assumptions of Corpus Linguistics to extract Specialized Unities of Translation (UTEs) from aforementioned Corpus using the Sketch Engine software. The most recurrent UTEs were organized in a databank with YAML records, and through a research system in Python, they are available for consultation. We hope, with this first step, to contribute to the creation of an agile and trustable support resource. Also, the system was made so that it can grow from the demand and availability of contributions, in order to support linguistically both interpreters and translators of the field, as well as the immigrants themselves.

Keywords: Multilingual Glossary. Corpus-driven Terminology. Terminological databank. Corpus Linguistics. Immigration and asylum.

RESUMEN

Cada año, miles de personas migran desde donde viven en busca de mejores oportunidades de vida. En 2018, 68,5 millones de personas se vieron obligadas a abandonar sus hogares; de ellas, 25,4 millones son refugiados y 3,1 son solicitantes de asilo. En Brasil, hemos batido el récord de solicitudes de asilo en 2017: 33.866. En este contexto, el grupo de investigación MOBILANG (Movilidades e idiomas en contacto) tiene como objetivo investigar los contactos lingüísticos que surgen de la movilidad humana y proporcionar soluciones lingüísticas para inmigrantes y refugiados en Brasil. Para ayudar a los intérpretes y traductores voluntarios que trabajan en este campo, una de las soluciones propuestas por el Grupo fue la creación de un material de referencia en línea sobre el tema. En este sentido, el objetivo del presente trabajo es establecer las bases necesarias para que el Glosario Multilingüe en Línea sobre Migración y Asilo estuviere disponible. Este glosario fue sistematizado a partir del Corpus Multilingüe sobre Migración y Asilo (COMMIRE, en portugués), que recopilamos a lo largo de tres años de iniciación científica. Desarrollado en colaboración con el grupo de investigación TermiTraDiCo (Terminología y traducción basada en el corpus), el glosario se basó en los supuestos teóricos de Lingüística de Corpus para extraer las Unidades Especializadas de Traducción (UTEs) del dicho corpus utilizando el software Sketch Engine. Las UTEs más recurrentes se organizaron en un banco de datos con registros YAML, y a través de un sistema de investigación en Python, están disponibles para consultación. Esperamos, con este primer paso, contribuir a la creación de un recurso de soporte ágil y confiable. Además, el sistema se creó para que pueda crecer a partir de la demanda y disponibilidad de contribuciones, con el fin de apoyar lingüísticamente tanto a los intérpretes y traductores del campo, como a los propios inmigrantes.

Palabras clave: Glosario multilingüe. Terminología dirigida por Corpus. Banco de datos terminológico. Lingüística de Corpus. Inmigración y asilo.

RÉSUMÉ

Chaque année, des milliers de personnes émigrent de leur lieu de vie à la recherche de meilleures opportunités. En 2018, 68,5 millions de personnes ont été forcées de quitter leur domicile, dont 25,4 millions sont des réfugiés et 3,1 millions des demandeurs d'asile. Dans ce contexte, le groupe de recherche MOBILANG (Mobilités et langues en contact) a pour objectif d'étudier les contacts linguistiques issus de la mobilité humaine et de fournir des solutions linguistiques aux immigrants et réfugiés au Brésil. Afin d'aider les interprètes et traducteurs bénévoles travaillant dans ce domaine, l'une des solutions proposées par le Groupe a été la création d'un document de référence en ligne sur ce sujet. En ce sens, l'objectif du présent travail est d'établir les bases nécessaires pour rendre disponible le Glossaire Multilingue En Ligne sur la Migration et L'asile. Ce glossaire a été systématisé à partir du Corpus Multilingue sur la Migration et L'asile (COMMIRE, en portugais), que nous avons compilé au cours de trois années d'initiation scientifique. Développé en partenariat avec le groupe de recherche TermiTraDiCo (Terminologie et Traduction Dirigée par Corpus), le Glossaire s'appuie sur les hypothèses théoriques de la Linguistique de Corpus pour extraire des Unités Spécialisées de Traduction (UTEs) dudit Corpus à l'aide du logiciel Sketch Engine. Les UTEs les plus récurrents ont été organisés dans une banque de données avec les enregistrements YAML et, grâce à un système de recherche en Python, ils sont disponibles pour consultation. Nous espérons, avec cette première étape, contribuer à la création d'une ressource de support agile et fiable. En outre, le système a été conçu de manière à ce qu'il puisse se développer à partir de la demande et de la disponibilité des contributions, afin d'aider linguistiquement les interprètes et les traducteurs du domaine, ainsi que les immigrants eux-mêmes.

Mots-clés : Glossaire multilingue. Terminologie dirigée pour corpus. Base de données terminologique. Linguistique de corpus. Immigration et asile.

LISTA DE GRÁFICOS

LISTA DE FIGURAS

LISTA DE TABELAS

ABREVIAÇÕES E SIGLAS

ACNUR – Alto Comissariado das Nações Unidas para Refugiados

CÁRITAS – Cáritas Brasileira

CONARE – Comitê Nacional para os Refugiados

COMMIRE – Corpus Multilíngue de Migração e Refúgio

CoMPLETT – Corpus Multilíngue para Pesquisas em Línguas Estrangeiras, Tradução e Terminologia

CSS – Cascade Style Sheets

DPU – Defensoria Pública da União

IMDH – Instituto de Migrações e Direitos Humanos

HTML – Hypertext Markup Language

MOBILANG – Mobilidades e Línguas em Contato

OI – Organização Internacional

OIM – Organização Internacional das Migrações

ONU – Organização das Nações Unidas

PF – Polícia Federal

SK – Sketch Engine

TermiTraDiCo – Terminologia e Tradução Direcionadas por Corpus

TCT – Teoria Comunicativa da Terminologia

TGT – Teoria Geral da Terminologia

UTE – Unidade de Tradução Especializada

YAML – YAML Ain’t a Markup Language

SUMÁRIO

1. INTRODUÇÃO

Desde os primórdios da civilização, o ser humano tem migrado constantemente, por motivos variados: de ordem econômica, cultural, religiosa, ou até mesmo por motivação pessoal. Atualmente, é muito comum que a imigração seja motivada pela busca de melhores condições de vida, novas oportunidades de emprego, de habitação – o que ocorre muito nos casos de migração do sul para o norte global.

A Organização Internacional para as Migrações (OIM), em seu relatório *Global Migration Indicators 2018*[1], estima que há cerca de 258 milhões de imigrantes no planeta atualmente. Desses, 68,5 milhões foram forçadas a deixar seu local de residência em razão de perseguição, conflito ou violência generalizada, conforme aponta o relatório *Global Trends – Forced Displacement in 2017*[2], da Agência da ONU para Refugiados (ACNUR). Ainda de acordo com o *Global Trends* (ACNUR, 2018), desses 68,5 milhões forçados a deixar suas residências, 25,4 milhões são refugiados e 3,1 milhões são solicitantes de refúgio. É importante destacar que 57% desses imigrantes vêm de três países: 6,3 milhões da Síria, 2,6 milhões do Afeganistão e 2,4 milhões do Sudão do Sul. Entre os países que mais recebem refugiados estão a Turquia (3,5 milhões), o Paquistão (1,4 milhão) e a Uganda (1,4 milhões).

Com o Brasil não poderia ser diferente. Segundo o relatório Refúgio em Números[3], produzido pelo CONARE (Comitê Nacional para os Refugiados), o Brasil concedeu refúgio a 10,145 refugiados entre 2010 até o fim de 2017. Além disso, 2017

[1] INTERNATIONAL ORGANIZATION FOR MIGRATION. Global Migration Indicators. 2018. Disponível em: http://publications.iom.int/system/files/pdf/global_migration_indicators_2018.pdf Acesso em: 03 jun 2019.
[2] UNHCR. Global Trends: Forced Displacement in 2017. 2018. Disponível em: https://www.unhcr.org/5b27be547#_ga=2.179681892.157476286.1560099262-257190238.1558393081. Acesso em: 03 jun 2019.
[3] ACNUR. Refúgio em Números. 3ª ed. 2018. Disponível em: https://www.acnur.org/portugues/wp-content/uploads/2018/04/refugio-em-numeros_1104.pdf. Acesso em: 03 jun 2019.

foi o ano em que batemos o recorde em número de solicitações de refúgio: 33.866 pedidos, mais que o dobro do ano anterior (10.308). É importante salientar que esses números não compreendem os venezuelanos e haitianos, que podem usufruir de visto humanitário e prerrogativas regionais diversas. Ainda assim, mais da metade das solicitações feitas em 2017 (17.865) vêm de venezuelanos. Em seguida, vêm os cubanos, com 2.373 solicitações, os haitianos, com 2.362 solicitações, e os angolanos, com 2.036 solicitações.

Esse contexto motivou a criação do grupo MOBILANG (Mobilidades e Línguas em Contato), cujo objeto de estudo principal é o contato de línguas proporcionado pelas mobilidades humanas. O grupo se subdivide em várias linhas de pesquisa e projetos de extensão com o objetivo de estudar esses contatos em suas diferentes perspectivas.

Estudos preliminares desenvolvidos pelo grupo ressaltam que a maior dificuldade enfrentada pelo imigrante ao chegar no Brasil é de ordem linguística (MIRANDA, 2016; MOLINA CABRERA, 2017; MILITÃO, 2017; GARCÍA, 2019). Com base nisso, uma solução mais imediata prevista pelos pesquisadores foi a implementação de um banco de intérpretes. Sabendo disso, as professoras Sabine Gorovitz, Susana Martínez e Carolina Capilla idealizaram e implementaram o projeto de extensão *Migrações e Fronteiras no Distrito Federal: a Integração Linguística como Garantia dos Direitos Humanos*, que se propõe a oferecer apoio linguístico aos solicitantes de refúgio em seu processo de solicitação junto ao CONARE. Esse apoio se dá por meio da disponibilização de intérpretes multilíngues voluntários, os quais comparecem ao CONARE em horário e dia marcados previamente para interpretar a entrevista de elegibilidade ao refúgio (GARCIA, 2019).

Outra solução cunhada pelo grupo foi a compilação de um banco de termos multilíngues que pudesse auxiliar tanto os intérpretes, tradutores e produtores de material

para os imigrantes, majoritariamente, quanto os próprios solicitantes de refúgio e refugiados. Para a execução de um projeto como esse, são necessários conhecimentos em Terminologia, Tradução e Linguística de Corpus. Por isso, o grupo MOBILANG fez uma parceria com o TermiTraDiCo (Terminologia e Tradução Direcionadas por Corpus), uma das linhas de pesquisa afiliadas ao grupo de pesquisa COMPLETT (Corpus Multilíngue para Pesquisas em Línguas Estrangeiras, Tradução e Terminologia), do qual faço parte, como aluna de Iniciação Científica sob a orientação da Profa. Dra. Elisa Duarte Teixeira. Este trabalho descreve, portanto, o processo de construção de um glossário multilíngue digital sobre refúgio e imigração que vimos desenvolvendo há quase três anos.

Conforme descreve o item seis do processo de solicitação de refúgio (na seção 2 deste trabalho), o solicitante passa por uma entrevista de elegibilidade na qual pode dispor de um intérprete, que auxilie na mediação da comunicação entre ele e o oficial de elegibilidade do CONARE. García (2019) pontua em seu trabalho que a tradução de um texto escrito é diferente da tradução em tempo real – interpretação – de uma entrevista, uma vez que a interpretação é efêmera, não há tempo para a pesquisa de qualidade e nem acesso em tempo real a qualquer material de apoio.

Além disso, com a intensificação da migração dos últimos anos, uma nova modalidade de intérprete surgiu: o intérprete comunitário, aquele que interpreta em contextos jurídicos, hospitalares, forenses, fazendo a mediação entre um serviço básico e a pessoa que necessita utilizá-lo, mas não domina a língua no qual o serviço é provido (ORIGUELLA, 2014 *apud* GARCÍA, 2019).

Pochkkacher (2004, p. 32 *apud* GARCÍA, 2019) destaca a necessidade de se publicar mais materiais voltados para intérpretes (e também tradutores), para que haja maior qualidade da interpretação. Nesse sentido, a produção de um glossário digital, como o que propomos produzir neste trabalho, facilita o acesso ao conteúdo especializado da

entrevista, de forma que o intérprete possa se preparar para a entrevista que vai mediar e, até mesmo, fazer consultas rápidas durante a interpretação.

No que se segue, começaremos por descrever, *en passant* e à guisa de justificativa, o processo de refúgio no Brasil, como se dá a solicitação de refúgio e qual é o papel do intérprete na entrevista de elegibilidade. Em seguida, relacionaremos a Tradução à Terminologia e abordaremos a Linguística de Corpus como abordagem teórica e metodológica de embasamento deste trabalho. Na sequência, apresentaremos dados sobre o corpus compilado para a extração dos termos e as fichas de armazenamento dos dados utilizadas. Por fim, descrevemos o processo de construção do banco de dados e da interface online do glossário.

2. O PROCESSO DE REFÚGIO NO BRASIL

No Brasil, o processo de refúgio é regulado pela Lei nº 9.474 de 22 de julho de 1997[4] (conhecida como Lei do Refúgio) e pela Lei nº 13.445 de 24 de maio de 2017[5] (conhecida como Lei da Migração).

É importante ressaltar que há diferenças entre a solicitação de refúgio e a solicitação de asilo político. A solicitação de asilo político se baseia no Estatuto de Roma do Tribunal Penal Internacional de 1998[6] e tem como objetivo fornecer proteção a uma pessoa que é perseguida por motivos políticos. Essa proteção é concedida somente depois que o processo de solicitação de asilo é deferido. Já a solicitação de refúgio é regulada pelo ACNUR e se apoia na Convenção de Genebra de 1951[7], e não se limita a prover proteção em casos restritos a perseguição política, mas prevê proteção também nos casos de perseguição por motivos de raça, religião, nacionalidade, grupo social ou opinião política[8]. A proteção garantida pelo refúgio vigora a partir do momento em que ele é solicitado pelo imigrante nas fronteiras, aeroportos ou portos.

De acordo com Militão (2017) e Jubilut (2014), o processo de solicitação de refúgio ocorre da seguinte maneira:

1. Em primeiro lugar, o imigrante deve fazer uma manifestação informal da vontade de solicitar refúgio nas fronteiras, aeroportos e portos na Polícia Federal (PF);

[4] PRESIDÊNCIA DA REPÚBLICA. Estatuto dos Refugiados. 1997. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9474.htm. Acesso em: 03 jun 2019.
[5] PRESIDÊNCIA DA REPÚBLICA. Lei de Migração. 2017. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/lei/l13445.htm. Acesso em: 03 jun 2019.
[6] PRESIDÊNCIA DA REPÚBLICA. Estatuto de Roma do Tribunal Penal Internacional. 2002. Disponível em: http://www.planalto.gov.br/ccivil_03/decreto/2002/D4388.htm. Acesso em: 03 jun 2019.
[7] ACNUR. Convenção de Genebra relativa ao Estatuto dos Refugiados. 1951. Disponível em: https://www.acnur.org/fileadmin/Documentos/portugues/BDL/Convencao_relativa_ao_Estatuto_dos_Refugiados.pdf . Acesso: 03 jun 2019.
[8] MINISTÉRIO DA JUSTIÇA. Entenda as diferenças entre refúgio e asilo. Disponível em: https://www.justica.gov.br/news/entenda-as-diferencas-entre-refugio-e-asilo. Acesso em: 03 jun 2019.

2. A PF repassa todo o material disponível (cartilhas, formulários e etc.) para o imigrante, a fim de esclarecer do que se trata o refúgio (nos termos do ACNUR) e como ocorre o processo no Brasil;

3. Caso o imigrante tenha entrado no país com o visto de turista, ele pode se dirigir à Cáritas Brasileira[9] e ao Instituto de Migrações e Direitos Humanos (IMDH) – ambas instituições não-governamentais católicas - para receber a acolhida e informações sobre o processo de solicitação de refúgio, sejam cartilhas, formulários e etc., isto é, o mesmo material disponibilizado nas fronteiras, aeroportos e portos, e também passar por uma rápida triagem para saber se o caso realmente se qualifica como solicitação de refúgio;

4. Depois de devidamente orientado, o imigrante dá início formal ao processo com o preenchimento do Termo de Declaração, por meio do qual se torna solicitante de refúgio. Este é o primeiro documento do solicitante até a expedição do Protocolo Provisório;

5. O solicitante deve comparecer à Caritas, IMDH ou PF para preencher um formulário com informações pessoais e solicitar a marcação da entrevista de elegibilidade. Nesse momento o Protocolo Provisório é emitido e o solicitante pode usufruir de direitos básicos a que todo brasileiro tem acesso (educação, trabalho, saúde, etc.);

6. O solicitante de refúgio é convocado para a entrevista de elegibilidade, que é confidencial e que pode ser mediada por um intérprete ou advogado.

[9] CÁRITAS BRASILEIRA. Quem Somos. Disponível em: http://caritas.org.br/quem-somos-e-historico. Acesso em: 03 jun 2019.

7. O conteúdo do formulário e da entrevista é armazenado, analisado pelos oficiais de elegibilidade do CONARE e um parecer é emitido, cuja análise é feita por outros oficiais de elegibilidade;

8. Em caso de deferimento, o status de refugiado é concedido ao solicitante e ele pode, então, solicitar a emissão do Registro Nacional do Estrangeiro (RNE) e permanecer em território nacional;

9. Em caso de indeferimento, ao solicitante é assegurado o direito de recurso, de fazer uma nova entrevista e de usufruir dos serviços da Defensoria Pública da União (DPU) para a elaboração desse recurso. Se, eventualmente, a decisão final for negativa, o solicitante fica em território nacional, mas deve se submeter às leis de migração locais. Em nenhuma hipótese o solicitante, tendo sua vida em risco, será enviado de volta ao seu país de origem.

Com base no exposto, é possível identificar que o imigrante terá contato com diversos materiais impressos (cartilhas, formulários, manuais e guias) que farão toda a diferença em sua compreensão sobre o processo de solicitação de refúgio, bem como no procedimento de solicitação em si. Como dissemos anteriormente, mais da metade dos migrantes do mundo vêm da Síria, Afeganistão e Sudão do Sul. Além disso, o Brasil bateu recorde de recebimento de solicitações de venezuelanos, haitianos, cubanos e angolanos. Sabe-se, também, que a língua tem influência direta na vida cotidiana das pessoas.

Na Síria, o árabe é a língua oficial, mas várias línguas minoritárias são faladas no dia a dia, como por exemplo, o árabe levantino do norte (conhecido também como árabe sírio-libanês), uma variação minoritária do árabe padrão, falado principalmente pelos

turcos na Síria[10]. Nesse mesmo país, há também a presença de línguas minoritárias como o adigue (também chamado de circassiano do oeste), o armênio, o azeri (ou azerbaijano), o curdo, e o turco[11]. Já o Afeganistão tem como línguas oficiais o pachto e o dari[12]. Os venezuelanos e cubanos tem como língua oficial o espanhol, ao passo que os haitianos têm o francês e o crioulo. Os angolanos têm como língua oficial o português e o chocué (ou quioco), mas o quicongo e ovambo são línguas nacionais e o umbundu e o quimbundo são línguas minoritárias amplamente faladas[13]. O Sudão do Sul fala inglês oficialmente, mas tem o árabe (crioulo sudanês) como língua nacional, e o bari e o zande são as línguas mais faladas regionalmente[14].

Estudos prévios de iniciação científica (FURTADO e GOROVITZ, 2017) mostram que os materiais fornecidos aos imigrantes e solicitantes de refúgio no Brasil são traduzidos, em sua maioria, para as línguas árabe, espanhola, inglesa e francesa. Além de não abarcarem várias das línguas oficiais e nacionais faladas pelos solicitantes, devemos salientar que o espanhol, o francês e o inglês podem ser, em muitos casos, a segunda ou terceira língua do imigrante, prejudicando assim seu processo de solicitação de refúgio (tanto na solicitação, como na entrevista). Daí, surge a necessidade de se produzir materiais de referência multilíngues de qualidade, tanto para a pessoa solicitante de refúgio e refugiada, quanto para o intérprete que irá fazer essa mediação.

[10] ETHNOLOGUE. **Syria**. Disponível: https://www.ethnologue.com/country/SY/status. Acesso em: 03 jun 2019.
[11] ETHNOLOGUE. **Syria**. Disponível: https://www.ethnologue.com/country/SY/status. Acesso em: 03 jun 2019.
[12] ETHNOLOGUE. **Afghanistan**. Disponível: https://www.ethnologue.com/country/AF/status. Acesso em: 03 jun 2019.
[13] ETHNOLOGUE**. Angola**. Disponível: https://www.ethnologue.com/country/AO/status. Acesso em: 03 jun 2019.
[14] ETHNOLOGUE. **South Sudan**. Disponível: https://www.ethnologue.com/country/SS/status. Acesso em: 03 jun 2019.

Ainda que haja no mercado um único dicionário sobre migrações, o Dicionário Crítico de Migrações[15], não há nenhum material que seja multilíngue, gratuito e online que possa de fato auxiliar os intérpretes, tradutores e refugiados a enfrentar o desafio linguístico das situações comunicativas que vão se deparar no preenchimento do formulário de solicitação de refúgio, por exemplo. Por isso, nosso objetivo é justamente preencher essa lacuna, produzindo um glossário multilíngue, em português, inglês, francês e espanhol, online que seja mais acessível. Para isso, utilizamos a Linguística de Corpus, exatamente porque essa abordagem se baseia no uso de textos reais, utilizados pelos imigrantes, refugiados e intérpretes no dia-a-dia, além de representar com maior fidelidade o público em questão. No próximo capítulo introduziremos toda a teoria que sustenta esse trabalho e depois partiremos para a explicação da metodologia utilizada.

[15] EDITORA UNB. **Dicionário Crítico de Migrações Internacionais**. 2017. Disponível em: https://loja.editora.unb.br/produto/856/dicionario-critico-de-migracoes-internacionais. Acesso em: 20 jun 2019.

3. REFERENCIAL TEÓRICO

Dado o contexto acima, agora, se faz necessário apresentar a teoria sobre a qual esse trabalho é embasado. Introduziremos a Lexicologia brevemente, devido à sua importância para a estruturação de trabalhos lexicográficos e terminográficos. Em seguida, discutiremos a Terminologia em sua interface com a Tradução Técnica, visto que este trabalho tem como público-alvo principal os tradutores e intérpretes da área de migração e refúgio e, por fim, apresentaremos a Linguística de Corpus, abordagem e metodologia sob a qual este trabalho se apoia.

3.1 LEXICOLOGIA, TERMINOLOGIA E TRADUÇÃO

A Lexicologia é a disciplina linguística que se debruça sobre o estudo das palavras de uma língua – o léxico – de forma integrada (LORENTE, 2004 *apud* GUERRA E ANDRADE 2012, p. 230). Na modernidade, esta disciplina passou a se basear nos conceitos linguísticos de Saussure (2006 [1916]), apresentados no século XX, como, por exemplo, a visão da língua como um sistema. Além disso, com as contribuições de Saussure e outros autores, a Lexicologia passou a ser vista como parte do plano funcional da língua, e não só como um conjunto de significados. A significação, estudada pela Semântica, deixou de ser uma exclusividade da Lexicologia (ORSI, 2012).

As investigações lexicológicas podem ser feitas com diversos focos, como, por exemplo, na semasiologia (definir conceitos, fazendo o percurso conceito → definição ou significado → significante) ou na onomasiologia (nomear conceitos, fazendo o percurso definição → conceito ou significante → significado); pode-se empreender, também, estudos diacrônicos (em várias épocas históricas) ou sincrônicos (em um dado recorte de tempo); diatópicos (diferentes regiões) ou sintópicos (numa região delimitada);

diastráticos (diferentes camadas sociais) ou sinstráticos (numa camada social específica); e, diafásicos (estilos diversos) ou sinfásicos (em um dado estilo) (ORSI, 2012).

Para Cabré (1999, p. 56), principal autora da Teoria Comunicativa da Terminologia (TCT), a língua é um sistema heterogêneo, complexo, composto de subsistemas inter-relacionados, os quais podem ser descritos em cinco níveis, que podem se complementar mutuamente (e formar novas unidades básicas de descrição linguística): fonológico (fonema), morfológico (morfema), lexical (lexema), sintático (sentença) e discursivo (texto). Para a autora, uma língua consiste de uma variedade de subcódigos que os falantes usam de acordo com suas necessidades e a natureza da situação comunicativa. Mesmo com essa variedade de subcódigos, as línguas possuem unidades e normas gerais internalizadas pelos falantes (idem, p. 59).

Ainda segundo Cabré (1999, p. 59), as línguas especializadas consistem em um conjunto de subcódigos que pode se sobrepor parcialmente aos subcódigos da língua geral. Os primeiros podem ser caracterizados por área específica do conhecimento, tipos de interlocutores, contexto e conteúdo específicos, etc. Em suma, as linguagens de especialidade são caracterizadas pelo uso de terminologia, ou seja, de termos que nomeiam conceitos de uma área de especialidade em questão (ANDRADE, 2001, p. 193). Assim como os lexemas da língua geral, os termos são unidades de significado específico que ocorrem em contextos específicos (BARROS, 2004, p. 40). E a ciência que se ocupa de seu estudo é a Terminologia.

De acordo com Krieger e Finatto (2004, p. 20), a Terminologia possui dupla identificação: primeiro "**T**erminologia", como uma disciplina da linguística que se encarrega do estudo dos termos, e "**t**erminologia", conjunto de palavras específicas de uma determinada área do conhecimento.

Os estudos em Terminologia iniciaram-se na Áustria, por volta de 1930, com o engenheiro Eugene Wüster, que tinha como principal objetivo padronizar o uso dos termos nas áreas de especialidade e eliminar acepções ambíguas, seguindo um percurso semasiológico e prescritivo – o que deu origem a uma teoria normativa da Terminologia batizada de Teoria Geral da Terminologia (TGT) (BARROS, 2004; KRIEGER e FINATTO, 2004).

Com o desenvolvimento da tecnologia nos anos 1960, os bancos de dados começaram a surgir e se difundiram, o que fez com que o caráter normatizador da Terminologia fosse internacionalizado. Entre os anos 1970 e 1990, a Terminologia se consolidou como disciplina científica, contribuindo para a normatização e expansão vocabular das línguas. Foi alvo de ações institucionais que consolidaram seu crescimento e permitiram o treinamento de profissionais da área (BARROS, 2004; KRIEGER e FINATTO, 2004).

Mas foi nos anos 1990 que a área teve seus pressupostos clássicos questionados e novas teorias foram propostas. Com a revisão da TGT, novos estudos terminológicos surgiram, em especial estudos descritivos, valorizando os aspectos comunicativos e sociais nas línguas de especialidade. É Gaudin quem, em 1993, propõe a Socioterminologia, voltada para o aspecto social das linguagens de especialidade, criticando os dicionários, glossários e materiais de referência que não representavam a variação que ocorre na realidade de uso cotidiano da língua (KRIEGER e FINATTO, 2004).

Depois disso, destaca-se o trabalho da professora Maria Teresa Cabré, que propõe a TCT, trazendo à luz o aspecto polissêmico e mutável dos termos, que sofrem variação e todo tipo de transformação pelo uso, assim como os demais vocábulos da língua geral (KRIEGER e FINATTO, 2004; CABRÉ, 1999).

A Terminologia tem três objetos de estudo principais: a) as unidades terminológicas; b) a definição terminológica e; c) os textos especializados (CABRÉ, 1999). Por estudar as linguagens de especialidade, a Terminologia permeia diversas áreas do conhecimento e é, portanto, interdisciplinar. E uma das áreas que possui relação íntima com a Terminologia é a Tradução, especialmente a Tradução Especializada, já que em sua própria origem e existência, são interdependentes. Nas palavras de Barros: "A tradução mantém também uma relação intrínseca com a Terminografia e com a Lexicografia, visto que estas últimas produzem um dos principais instrumentos de trabalho do tradutor: os dicionários." (BARROS, 2004, p.79).

Ainda que essa relação seja tão evidente, como apontam vários autores, a confecção de dicionários e glossários é tarefa de difícil empreendimento, e que nem sempre tem como público-alvo principal o tradutor, como mostra Teixeira (2008). Mesmo que a comunicação especializada seja levada em conta para a produção terminográfica, segundo as teorias mais modernas, o termo, o conceito e a definição são sempre os focos.

Todavia, ainda que um tradutor/intérprete saiba a terminologia correta da área de especialidade, não há garantia de que sua tradução será satisfatória. É preciso que conheça também as características do gênero e do tipo textual com que está trabalhando nas duas línguas, e também como a língua geral se combina aos termos nos textos técnicos daquele gênero e tipologia textuais (TEIXEIRA 2008, p. 6-11). E é nesse ponto que abordagem proposta pela Linguística de Corpus se torna fundamental para a elaboração do glossário resultante desta pesquisa.

As unidades de sentido contidas em um discurso especializado escrito ou falado, a ser traduzido para uma outra língua, nem sempre se limitam ou até mesmo contêm um termo, conforme pontua Teixeira (2008, p. 11). Seguindo a proposta teórica e metodológica desta autora, chamaremos essas unidades de Unidades de Tradução

Especializadas (UTEs), pois, a nosso ver, o conceito de UTE é o que melhor representa a noção de unidade de sentido nos contextos especializados da perspectiva da tradução/interpretação, como é o caso da área da imigração e refúgio que enfocamos neste trabalho.

3.2 LINGUÍSTICA DE CORPUS

A Linguística de Corpus (LC) é uma abordagem empírica de estudo da língua que parte da análise de uma grande quantidade de textos em formato eletrônico – os corpora (TEIXEIRA, 2008, p. 368; TAGNIN, 2015, p. 19; BERBER SARDINHA, 2004, p. 30). Berber Sardinha, um dos professores que introduziu a Linguística de Corpus no Brasil, oferece em seu livro de 2004, debutante na área no país, conceitos essenciais para a compreensão dessa abordagem. Uma ideia central para a LC é que a língua é entendida como um sistema probabilístico, seguindo a Linguística Probabilística de Halliday (HALLIDAY, 1961 *apud* BERBER SARDINHA, p. 30), ou seja, "embora muitas construções sejam possíveis, algumas delas têm probabilidade maior de ocorrer" (TAGNIN, 2015, p. 20).

Partindo dessa visão, Biber (1998 *apud* BERBER SARDINHA 2004, p. 31) constatou que "há uma correlação entre características linguísticas e situacionais (contextos de uso)", isto é, a linguagem obedece a padrões, os quais são percebidos pela recorrência.

Essa visão probabilística da linguagem teve seu primeiro marco teórico por volta de 1960, com o corpus SEU (Survey of English Language), ainda em fichas (e não eletrônico), o qual estabeleceu o padrão para a compilação de corpora futuros. Em

seguida, surgiu o Corpus Brown, o primeiro informatizado, ainda na década de 1960, usando como exemplo o Corpus SEU (McENERY e HARDIE, 2012).

Contudo, a Linguística de Corpus evoluiu mesmo a partir dos anos 1980, com a popularização dos computadores. Isso permitiu a criação, mais tarde, do dicionário COBUILD[16], fruto da parceria entre a Universidade de Birmingham e a Editora Collins e o primeiro elaborado com base em corpora eletrônico (McENERY e HARDIE, 2012, p. 80; TEIXEIRA, 2008, p. 152). Atualmente, muitos dos dicionários e glossários disponíveis em língua inglesa foram compilados com ferramentas da LC, tais como o Dicionário Cambridge[17], o Dicionário Oxford[18] e o Dicionário Macmillan[19], o Dicionário Le Petit Robert[20] do francês, entre outros.

Bowker e Pearson (2002) e Teixeira (2008) oferecem em seus trabalhos critérios-guia para o planejamento e coleta de um corpus para que este possa ser usado em pesquisas de LC:

a) os textos devem ser **autênticos**, ou seja, não devem ter sido produzidos para estudo linguístico;

b) os textos devem ser **naturais**, produzidos por falantes nativos, ou deve-se explicitar quando não forem;

[16] COLLINS. English Dictionary. Disponível em: https://www.collinsdictionary.com/dictionary/english. Acesso em: 03 jun 2019.
[17] CAMBRIDGE DICTIONARY. Dicionário Cambridge. Disponível em: https://dictionary.cambridge.org/pt/. Acesso em: 03 jun 2019.
[18] OXFORD DICTIONARIES. Welcome to Oxford Dictionaries. Disponível em: https://languages.oup.com. Acesso em: 03 jun 2019.
[19] MACMILLAN DICTIONARIES. Macmillan Dictionary. Disponível em: https://www.macmillandictionary.com. Acesso em: 03 jun 2019.
[20] LE PETIT ROBERT. Le Dictionnarie. Disponível em: https://www.le-dictionnaire.com. Acesso em: 03 jun 2019.

c) os textos devem ser **representativos** da área à qual pertencem, e na maioria das vezes, quanto mais textos, melhor;

d) os textos devem ser **eletrônicos** para que possam ser processados pelo computador;

e) os textos devem ser escolhidos com **propósito**, isto é, deve-se limitar quais textos serão coletados e o porquê.

Segundo Berber Sardinha (2004, p. 90), os trabalhos em LC observam, em essência, os seguintes fenômenos:

a) **ocorrência** – só se pode observar o que foi mostrado pelos dados;

b) **recorrência** – é necessário que um fenômeno ocorra com regularidade para ser observado;

c) **coocorrência** – uma ocorrência deve ser relacionada a outros fatores linguísticos para que generalizações possam ser feitas.

Ainda de acordo com Sardinha (2004, p. 40), os resultados devem ser estudados à luz de três conceitos-chave:

a) **colocação**: relação entre itens lexicais ou entre itens do léxico e do campo semântico;

b) **coligação**: associação entre itens gramaticais e lexicais, e;

c) **prosódia semântica**: associação entre itens lexicais e significados específicos ou conotação (positiva, neutra ou negativa).

Na seção 4.2, apresentaremos as ferramentas de LC utilizadas para a extração dos dados linguísticos que compõem este trabalho. A seguir, discutiremos brevemente a relação entre LC, Tradução e Terminologia.

3.3 Terminologia direcionada por Corpus

Os estudos que associam a Linguística ou a Terminologia à Linguística de Corpus podem ser de dois tipos: baseados em corpus (*corpus-based*) ou direcionados por corpus (*corpus-driven*), segundo Tagnin (2009, p. 1084) e McENERY e Hardie (2012, p. 5-6).

Alguns autores preconizam que as pesquisas **baseadas em corpus** geralmente utilizam o corpus como fonte de dados para testar ou avançar uma teoria ou hipótese linguística pré-definida, geralmente consolidada na área, com vistas ao seu refinamento, validação ou refutação (McENERY e HARDIE, 2012, p. 5-6). Nesse sentido, pode-se dizer que utilizam a LC como metodologia (TAGNIN, 2009, p. 1084; McENERY e HARDIE, 2012, p. 6).

Por outro lado, as investigações **dirigidas por corpus** usam os dados levantados no corpus por meio das ferramentas criadas especificamente para esse fim como fonte primordial de observação e teorização de fenômenos linguísticos. Neste caso, a LC é tida não mais como metodologia, mas como abordagem, como forma de enxergar a língua (TAGNIN, 2009, p. 1084; McENERY e HARDIE, 2012, p. 6).

Este trabalho tem por objetivo extrair, de uma perspectiva multilíngue e com vistas à tradução, o conjunto de palavras e combinatórias de palavras recorrentes em textos da linguagem especializada da imigração, mais especificamente as solicitações de refúgio – isto é, as Unidades de Tradução Especializadas da área – à luz da LC. Esta será

usada, portanto, como abordagem e como metodologia, direcionado por corpus, para compilar um banco de dados inicial e apresenta-los numa interface (glossário) online.

3.4 Bancos de dados terminológicos

Para que se possa compilar um glossário consistente, é crucial que se tenha um banco de dados para o armazenamento das informações linguísticas extraídas do corpus. Para Krieger e Finatto (2004, p. 145), "um banco de dados terminológico apresenta-se como um sistema de informações interconectadas. Armazenado em um computador, visa a atender as necessidades de consulta de um grupo definidos de usuários". As autoras apontam ainda que há três características básicas inerentes às características de arquitetura de bancos de dados terminológicos: integração, estruturação e volume de informações (KRIEGER e FINATTO, 2004, p. 146).

A integração diz respeito à multiplicidade de fontes de que se originam os dados linguísticos a serem oferecidos em uma única plataforma de acesso. Essas fontes podem ser glossários, dicionários, vocabulários, ou, no caso deste trabalho, um corpus (KRIEGER e FINATTO 2004, p. 146). No que tange à estruturação, ela se faz inicialmente por meio de fichas terminológicas ou lexicológicas, as quais chamaremos aqui de fichas de UTEs. Essas fichas registram os dados das UTEs levantadas no corpus em campos pré-definidos (KRIEGER e FINATTO, 2004, p. 147). Na seção 4.2.2., mostraremos os campos que as compõem e daremos exemplos de fichas de UTEs preenchidas.

Por fim, o volume de informação é extremamente importante quando montamos um banco de dados. Krieger e Finatto (2004, p. 147) chamam a atenção para essa necessidade. Como estamos fazendo um trabalho que dispõe de pouco tempo e é de difícil

empreendimento, poderíamos dizer que o que vamos apresentar é um mini banco de dados temático, pois faz alusão ao nosso recorte temático, imigração e refúgio, e visa ser o ponto de partida para a construção de um banco de dados em constante ampliação, no âmbito dos projetos de que participa.

Quanto à parte operacional de se criar tal banco de dados, optamos por escrever o código em *Python*, utilizando fichas em *YAML*, no ambiente de programação *App Engine* do Google. Para que os usuários possam acessar os dados, produziremos uma interface com Python e HTML. Explicaremos o processo de construção do banco de dados e da interface de acesso nos capítulos 5 e 6 deste trabalho.

3.5 Glossários online

Atualmente, recebemos e transmitimos informações o tempo todo por meio de mídias sociais, e-mails, blogues, aplicativos de conversa, entre outros. O fim do século XX marca o surgimento da Sociedade de Informação, definida por Castells (2003) como uma sociedade em rede e fundada no poder proporcionado pela informação. Essa sociedade trouxe consigo a revolução promovida pelas tecnologias da informação, que transformaram totalmente a maneira como nos comunicamos diariamente e, especialmente, como geramos e difundimos a informação (CASTELLS, 2003). Em vista disso, optamos por hospedar e disponibilizar o glossário alvo desta pesquisa online, de forma que se tornasse mais acessível e útil aos usuários-alvo. Além disso, Teixeira (2008, p. 324) elenca diversas vantagens de se veicular um dicionário em ambiente virtual, em vez de impresso: consulta ágil, aspecto visual mais trabalhado, busca refinada e limitação espacial praticamente inexistente.

No que concerne os glossários especializados, de acordo com Barros (2004, p. 133), podem ser entendidos como obras terminográficas que compreendem um conjunto de palavras de uma área especializada. Ainda para essa autora, toda obra terminográfica ou lexicográfica pode ser genericamente chamada de dicionário ou repertório. Quando se fala na conceituação das obras lexicográficas e terminográficas, há uma discussão bastante acalorada e muitos autores divergem entre si, como mostra o artigo de Barbosa (2001). Neste trabalho, usaremos a definição proposta por Barros:

> **Glossário** (termo tolerado: dicionário bilíngue, dicionário multilíngue): pode situar-se tanto no nível do sistema como no da(s) norma(s). Sua principal característica é não apresentar definições, mas tão somente uma lista de unidades lexicais ou terminológicas acompanhadas de seus equivalentes em outras línguas. (BARROS 2004, p. 144)

Dito isso, Barros (2004, p. 151) define ainda em seu livro partes importantes de um repertório: a macroestrutura e a microestrutura. Em suas palavras, "por macroestrutura entende-se a organização de uma obra lexicográfica ou terminográfica". Ainda segundo a autora, essa organização se refere ao modo de organização dos verbetes, à presença de índices, imagens, mapas de conceitos, à apresentação da obra, entre outros. Já a microestrutura está relacionada à organização das informações apresentadas nos verbetes. Segundo a mesma autora, é necessário considerar o número de informações que serão apresentadas ao usuário, a presença dessas informações em todos os verbetes e a maneira com que essas informações serão ordenadas.

De acordo com Barbosa (1990, *apud* Barros 2004, p. 157), os verbetes são organizados em três microparadigmas, que são:

a) **paradigma informacional –** faz alusão à classificação morfológica, ao gênero, à conjugação (no caso dos verbos), à pronúncia etc.;

b) **paradigma definicional –** está relacionado a quantidade de semas que o verbete carrega;

c) **paradigma pragmático –** está relacionado aos contextos em que o verbete ocorre.

O paradigma definicional, conforme dissemos anteriormente, apoiadas na definição de "glossário" de Barros (2004, P. 144), não fará parte de nossa microestrutura, e por isso não trataremos da mesma neste trabalho.

No capítulo 6 apresentamos o modelo de macro e microestrutura que utilizamos para a composição do nosso glossário. Agora, passamos ao capítulo de metodologia, onde forneceremos informações sobre a compilação do corpus em que coletamos os dados, as ferramentas e metodologia de extração utilizadas e a estruturação interna do banco de dados e das fichas que abrigaram as informações terminológicas coletadas no corpus.

4. METODOLOGIA

4.1 A coleta do Corpus COMMIRE

Na seção 3.2, expusemos os critérios defendidos por Bowker e Pearson e por Teixeira para o planejamento e coleta de um corpus especializado. A coleta do corpus utilizado neste projeto – a que demos o nome de COMMIRE - **Co**rpus **M**ultilíngue de **Mi**gração e **Re**fúgio – foi feita ao longo de três anos, iniciada por meio de um trabalho de iniciação científica feito por mim e orientado pela professora Dra. Sabine Gorovitz (FURTADO e GOROVITZ, 2017), onde sistematizamos a metodologia de coleta de um corpus multilíngue sobre imigração e refúgio. Este trabalho foi sucedido por um segundo projeto de iniciação científica, também feito por mim, e orientado pela professora Dra. Elisa Duarte Teixeira. Neste projeto, que será concluído em julho de 2019, iniciamos a coleta de um corpus multilíngue comparável sobre imigração e refúgio, etapa necessária para a compilação de um banco terminológico. Nesta seção, apresentamos primeiramente os critérios para a coleta empregados, a metodologia de preparo dos textos e as ferramentas e rotinas empregadas na extração dos dados e preenchimento das fichas. Por fim, passamos para a análise dos dados levantados no corpus.

4.1.1. Critérios e procedimento de coleta do corpus COMMIRE

Antes de iniciar a coleta de dados em si, foi necessário estabelecer os parâmetros que guiariam o processo de seleção dos arquivos. A tabela a seguir sintetiza os critérios utilizados para o planejamento da coleta do corpus COMMIRE.

Tabela 1 - Critérios utilizados no planejamento do Corpus COMMIRE.

Critério	Características
Língua	Multilíngue (em português, espanhol, francês e inglês).
Tipo de Corpus	Comparável (maioria de textos originalmente escritos em cada uma das línguas; entretanto, traduções serão bem-vindas desde que coletadas de fontes confiáveis e que descrevam a procedência do material).
Tamanho	Médio-grande (pelo menos 1 milhão de palavras em cada língua, totalizando 4 milhões de palavras).
Modo	Textos escritos para refugiados, solicitantes de refúgio e imigrantes, ou para pessoas que trabalham com esses públicos.
Fontes	Governo, Organizações Internacionais (OIs) ou ONGs (Organizações não-Governamentais) que recebem e/ou trabalham com esse público.
Domínio do conhecimento	Refúgio e imigração.
Autoria	Produzidos e/ou publicados pelas instituições definidas no campo Fontes.
Data de Publicação	Recente, não mais que 15 anos.
Codificação	Todos os arquivos serão convertidos para o formato de texto sem formatação (*.txt*) em UTF-8, será acrescentado um cabeçalho constando dados bibliográficos e será feita a etiquetagem morfossintática automática do programa no programa de exploração de arquivos escolhido para a análise.

Fonte: elaboração própria.

Uma vez planejada a coleta, passamos a localizar os materiais disponíveis para o solicitante de refúgio e para o imigrante na internet. Esses materiais deveriam informar como se dá o processo de refúgio, como solicitar refúgio, quais são os documentos necessários para tal, entre outros. Portanto, escolhemos palavras-chaves para a busca de materiais nas línguas-alvo, a saber: “como solicitar refúgio”, “comment demander asile”, “how to claim asylum”, “solicitud de refugio”, “material para solicitação de refúgio”,

“guidance on claiming asylum”, “conseils sur la demande d'asile” e “orientación sobre la solicitud de asilo”.

Fomos coletando o material em cada língua na ordem em que aparecia. Ainda que o Google tenha seus próprios critérios de ranqueamento[21] de páginas, sabe-se que ele privilegia o conteúdo que é mais frequentemente acessado, o que justamente era o nosso interesse.

Uma das constatações que fizemos nessa etapa é que há muito mais material para quem trabalha na acolhida do solicitante de refúgio e imigrante do que para o solicitante de refúgio em si. Isso se deve, provavelmente, à necessidade de treinamento e orientação de pessoal, devido à natureza sensível da condição dos refugiados e imigrantes.

Cabe frisar que todo o processo de coleta foi bastante penoso, pois os materiais foram selecionados manualmente e passaram por análise cuidadosa das fontes, para ver se eram confiáveis, se o material foi produzido de forma responsável, se foi revisado ou não, se a autoria do material era múltipla e se foi produzido por falantes nativos, multilíngues ou traduzido.

Como determinamos o parâmetro de pelo menos milhão de palavras por língua, demoramos muito tempo nessa etapa, e, por causa disso, não foi possível transformar todo o corpus, que estava em *.pdf*, para *.txt* para esta pesquisa. Entretanto, os arquivos transformados passaram pelo *Abbyy Fine Reader*[22] (daqui em diante Abbyy), um programa pago de reconhecimento óptico de caracteres (OCR). Os arquivos resultantes

[21] GOOGLE. Como funciona a pesquisa. Disponível em: https://www.google.com/search/howsearchworks/. Acesso em: 27 jun 2019.
[22] ABBYY FINE READER. Características. Disponível em : https://www.abbyy.com/pt-br/finereader/in-details/. Acesso em: 27 jun 2019.

foram convertidos para o novo formato *.txt*, em codificação utf-8, e receberam um cabeçalho no qual constam informações bibliográficas do documento.

Como a tarefa de coleta demandou muito tempo, a maioria dos arquivos foi inserida em *.pdf* no programa de exploração, *Sketch Engine* (o qual será apresentado na seção 4.2.1.). O programa faz seu OCR, mas sabe-se que, assim como os arquivos transformados pelo Abbyy, todos necessitam de tratamento manual, devido à quebra de palavras e troca de elementos no processo de reconhecimento óptico. Infelizmente, não foi possível dar a todos os arquivos o tratamento adequado (identificar e consertar manualmente todos os problemas de conversão) – isso foi o que, provavelmente, causou muito ruído no corpus (como palavras quebradas, presença de símbolos, entre outros).

É importante ressaltar que todos os arquivos, até mesmo as revistas (de cunho informativo, produzido por especialistas da área), ainda que tenham sido coletados de um único organismo, tem autoria distinta e diversa. Todos os materiais estão disponíveis na internet de forma gratuita e foram produzidos institucional e coletivamente, seja pelo governo dos países em questão, seja por Organizações Internacionais (com o ACNUR ou a UNHCR), ou por ONGs que se responsabilizam pela acolhida dos refugiados, solicitantes de refúgio ou imigrantes. A maioria dos materiais possui autoria não identificada, pois foram produzidos de forma conjunta.

A área de imigração e refúgio é bem especializada, e como tal, não possui oferta equitativa de material em todas as línguas. Por isso, em muitos casos há disparidade na quantidade de palavras dos materiais coletados referentes a cada tipologia textual, na tentativa de balancear o corpus.

4.2. Ferramentas e rotinas de exploração do corpus e extração de UTEs

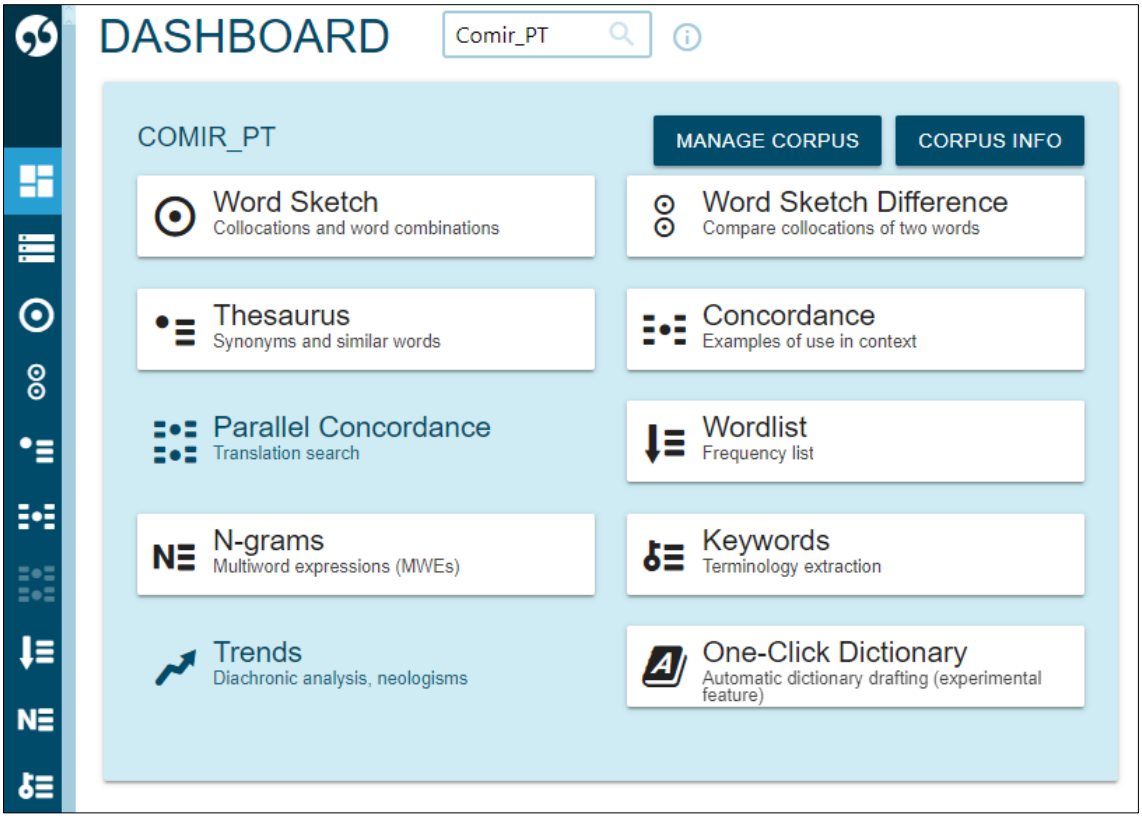
O programa que escolhemos para a exploração e extração de UTEs do corpus se chama *Sketch Engine*[23] (KILGARRIFF et. al., 2004), doravante SK, um programa de gerenciamento e análise de corpora desenvolvido pela *Lexical Computing Ltda.* desde 2003. Ele permite que, uma vez feito o upload dos textos, o seu corpus possa ser acessado em qualquer computador conectado à internet. O que determinou a escolha desse programa foi o fato de ele lidar bem com corpora multilíngues, possuir diversos corpora multilíngues de tamanho grande acoplados, que puderam ser utilizados para comparação, além de ferramentas que serão discutidas mais abaixo, como o *Word Sketches*. Aqui, é importante mencionar que se trata de um programa pago, que conta com um mês de teste para o *upload* de um corpus de até um milhão de palavras. Como extrapolamos esse tamanho, tivemos de pagar. Passaremos agora para a descrição das ferramentas utilizadas.

4.2.1 Ferramentas de LC no Sketch Engine

O SK conta com diversas ferramentas para a exploração de corpora. Depois que o upload do corpus é feito, a seguinte tela é exibida:

[23] SKETCH ENGINE. What is Sketch Engine. Disponível em: https://www.sketchengine.eu/#blue. Acesso em: 25 jun 2019.

Figura 1 - Captura da tela inicial do Sketch Engine



Fonte: elaboração própria.

Na Figura 1, podemos observar a presença das seguintes funcionalidades:

a) **Word Sketches** – ferramenta que sumariza o comportamento lexical e gramatical de uma expressão de busca;

b) **Word Sketch Difference** – oferece uma comparação do comportamento lexical e gramatical entre duas expressões de busca;

c) **Concordance** – mostra exemplos da expressão de busca em contexto;

d) **Thesaurus** – oferece sinônimos e expressões similares à expressão de busca;

e) **Parallel Concordance** – busca uma expressão em uma língua de partida e retorna o contexto da expressão de partida alinhado ao contexto da expressão de chegada (só funciona em corpora paralelos alinhados[24]);

f) **Wordlist** – ordena as palavras do corpus de acordo com a opção selecionada (frequência, alfabeticamente, etc);

g) **N-grams** – produz listas de frequências de uma sequência de palavras;

h) **Keywords** – extrai palavras-chave (obtidas pela comparação com a Wordlist de outro corpus);

i) **Trends** – análise diacrônica de um corpus;

j) **OneClick Dictionary** – ferramenta experimental de design de dicionários.

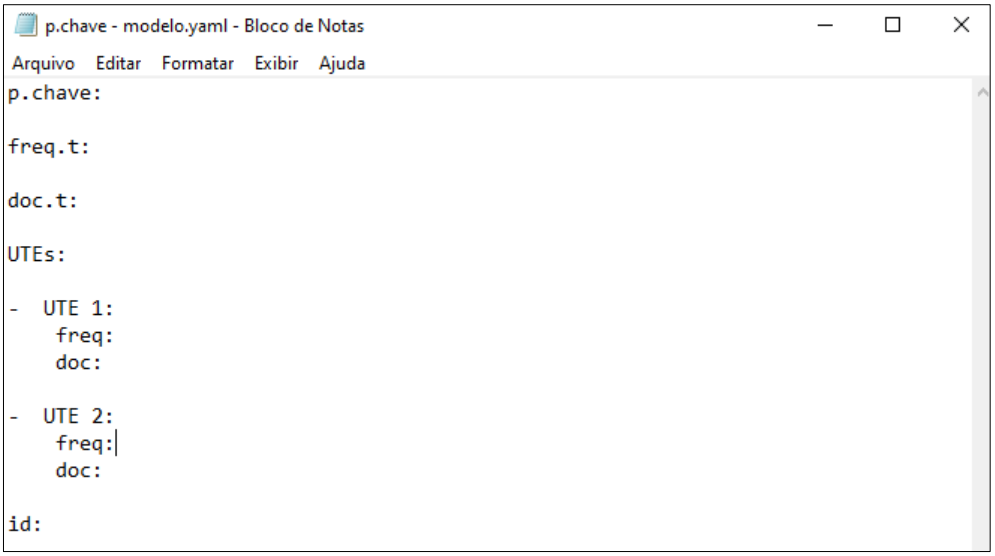
Neste trabalho, utilizamos as seguintes ferramentas para a extração dos dados: Word Sketches, Concordance e Keywords, as quais descreveremos a fundo mais a frente. Vejamos agora a estrutura da ficha de UTE e como as informações foram extraídas.

4.2.2 O modelo de ficha para coleta de UTEs

Levando em consideração as necessidades do nosso público-alvo, definimos dois modelos de ficha padrão para as UTEs. O primeiro consiste no que nomeamos de **ficha-mãe**, a qual contém os dados da palavra-chave extraída por meio da ferramenta Keyword. O segundo traz os dados das UTEs resultantes, que chamamos de **ficha-filha** – usadas para registrar todas as UTEs recorrentes que possuem pelo menos uma palavra-chave.

[24] Originais alinhados, no nível da sentença, com sua(s) respectiva(s) tradução(ões).

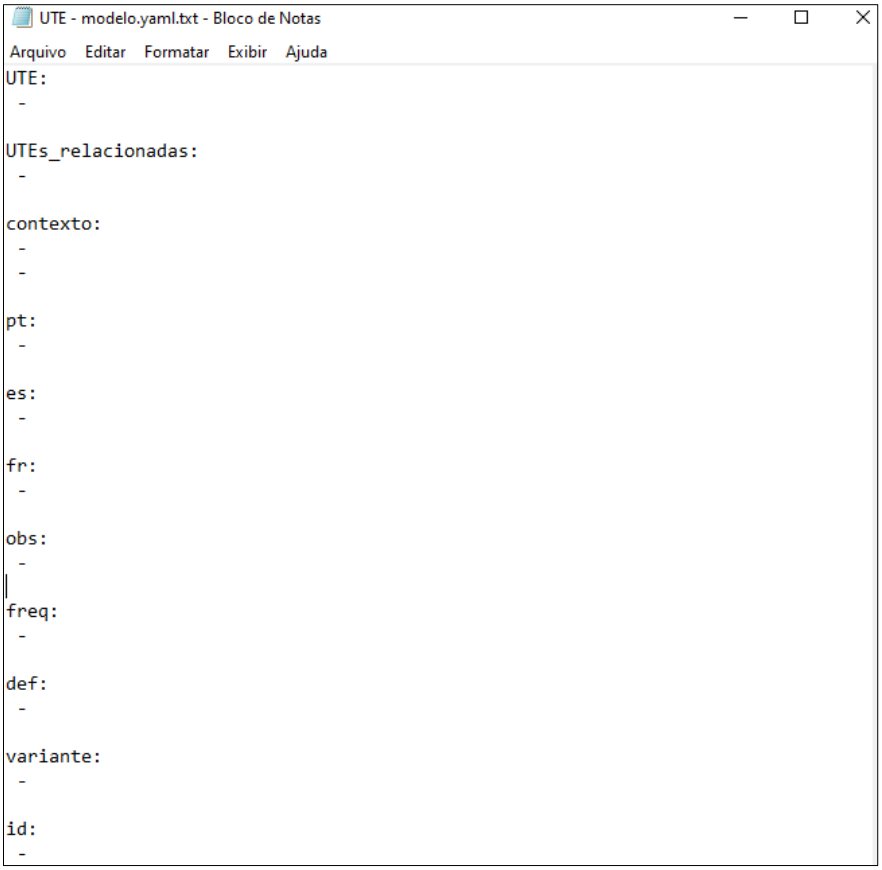
Figura 2 - Captura de tela do modelo da Ficha-mãe



Fonte: elaboração própria.

O primeiro campo, **p.chave**, consiste na palavra-chave obtida por meio da extração da lista de palavras-chave. Em segundo lugar, temos **freq.t**, que significa frequência total – é a frequência total daquela palavra-chave obtida por meio da lista de palavras-chave. Em seguida, há **doc.t**, que é a quantidade total de arquivos nos quais a palavra-chave ocorre. Sequencialmente, elencamos as UTEs que gerarão, cada uma, uma ficha-filha, com seus dados. Cada campo de UTE na Figura 2 apresenta os campos **freq** e **doc** que correspondem à frequência daquela UTE no corpus e em quantos documentos ela ocorre, respectivamente. Por fim, há o campo **id** que numera a ficha para nosso controle interno.

Figura 3 - Captura de tela do modelo da ficha de UTE



Fonte: elaboração própria

A Figura acima corresponde a um modelo de ficha de UTE. Nela, podemos observar os seguintes campos:

- UTE – corresponde a UTE em si;
- UTEs_relacionadas – elenca UTEs que possuem núcleo semelhante;
- Contexto – apresenta duas sentenças extraídas da ferramenta Concordance;
- PT – apresenta a UTE equivalente em português (se ocorrer no corpus português);
- ES – apresenta a UTE equivalente em espanhol (se ocorrer no corpus espanhol);

- FR – apresenta a UTE equivalente em francês (se ocorrer no corpus);
- EN[25] – apresenta a UTE equivalente em inglês (se ocorrer no corpus);
- OBS – observações pertinentes coletadas por meio das ferramentas;
- FREQ – frequência relativa da UTE no corpus (frequência da UTE dividida pela frequência da palavra-chave multiplicado por 100 e expressa com duas casas decimais de precisão);
- DEF – campo de definição da UTE, o qual não será preenchido no empreendimento deste trabalho, mas consta para inclusão posterior;
- Variante – campo destinado às siglas ISO correspondentes ao país dos arquivos nos quais a UTE ocorre;
- ID – numeração da ficha para controle interno.

4.3 Rotina de coleta de dados e preenchimento das fichas

4.3.1. Coleta de palavras-chave no Sketch Engine

Com a ferramenta Keywords é possível levantar as palavras-chave de um corpus – aquelas cuja frequência relativa no corpus de estudo é significativa, na comparação com um corpus de referência – e candidatos a termo[26], que são expressões multi-palavra mais frequentes no corpus de estudo se obtidas em comparação a um corpus de referência, além de corresponderem a estruturas terminológicas típicas da língua. As palavras-chave são ordenadas por ordem decrescente de chavicidade – que consiste em uma variável obtida por meio de um método chamado *simple maths*[27], no qual o programa normaliza a

[25] Não exibida na Figura acima pois corresponde à língua da ficha.
[26] SKETCH ENGINE. Keywords and term extraction. Disponível em: https://www.sketchengine.eu/guide/keywords-and-term-extraction/#toggle-id-2. Acesso em: 27 jun 2019.
[27] "Simple maths is a method for identifying keywords of one corpus vs another. It includes a variable which allows the user to turn the focus either on higher, or lower frequency words". (KILGARIFF, 2009).

frequência (relativiza e a calcula por milhão) no corpus de estudo e no corpus de referência.

Os corpora de referência usados para fins de extração de palavras-chave e candidatos a termo foram os seguintes:

Tabela 2 - Relação e características dos corpora de referência

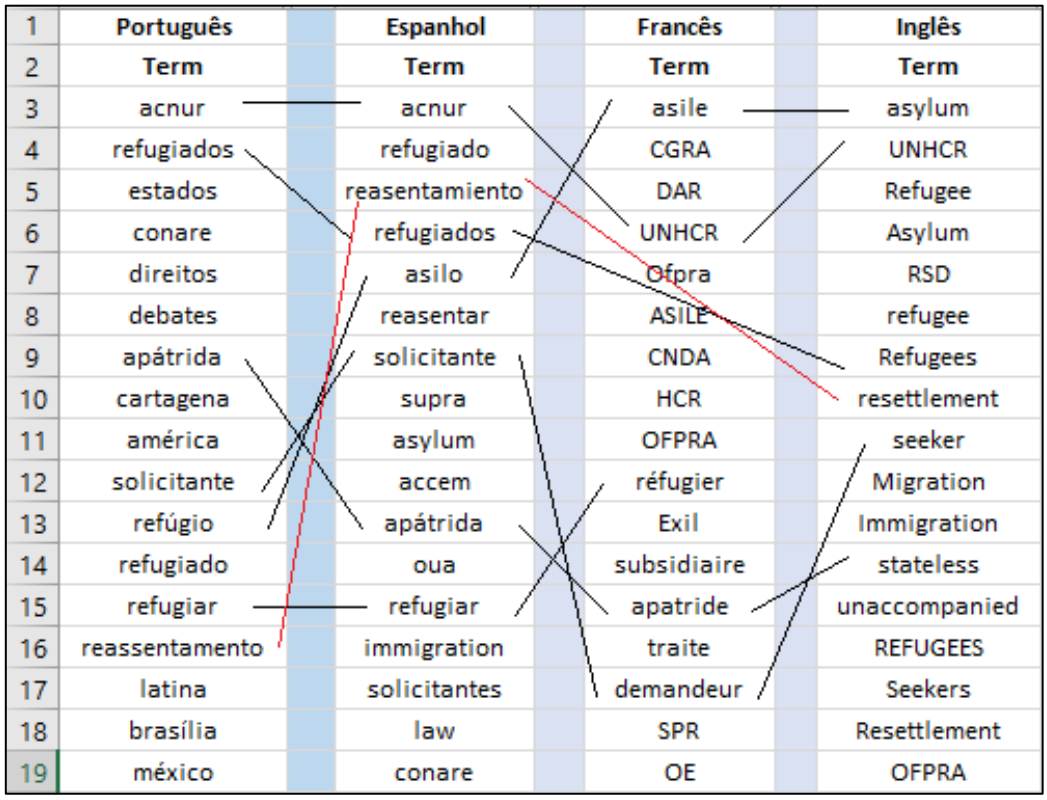
Idioma	Corpus	Quantidade de palavras	Variedades contidas
Inglês	English Web corpus 2015 (enTenTen15[28])	15 bilhões	Não há detalhes
Espanhol	Spanish Web corpus 2011 (esTenTen11[29])	9.5 bilhões	Espanhol europeu e americano
Francês	French Web corpus 2012 (frTenTen12)[30]	10 bilhões	Francês africano, europeu e canadense
Português	Corpus Brasileiro	7.7 bilhões	Português brasileiro

Fonte: elaboração própria

Para coletar as palavras-chave, escolhemos os corpora acima e selecionamos a frequência mínima de três ocorrências no corpus para listagem. O programa dá a opção de fazer o download dos arquivos em diversos formatos, e por isso optamos por utilizar em *.xls* (arquivo de tabela do *Excel*). Uma vez coletadas todas as palavras-chave neste formato, colocamos as listas lado a lado para identificação manual de equivalências *prima-facie*, como mostra a figura abaixo:

[28] SKETCH ENGINE. EnTenTen English Corpus. Disponível em: https://www.sketchengine.eu/ententen-english-corpus/#toggle-id-3 .Acesso em: 27 jun 2019.
[29] SKETCH ENGINE. EsTenTen Spanish Corpus. Disponível em: https://www.sketchengine.co.uk/esTenTen-spanish-corpus . Acesso em: 27 jun 2019.
[30] SKETCH ENGINE. FrTenTen French Corpus. Disponível em: https://www.sketchengine.eu/frtenten-french-corpus/. Acesso em: 27 jun 2019.

Figura 4 - Captura de tela da representação do modelo de identificação de equivalências entre palavras-chave.



1	Português		Espanhol		Francês		Inglês
2	Term		Term		Term		Term
3	acnur		acnur		asile		asylum
4	refugiados		refugiado		CGRA		UNHCR
5	estados		reasentamiento		DAR		Refugee
6	conare		refugiados		UNHCR		Asylum
7	direitos		asilo		Ofpra		RSD
8	debates		reasentar		ASILE		refugee
9	apátrida		solicitante		CNDA		Refugees
10	cartagena		supra		HCR		resettlement
11	américa		asylum		OFPRA		seeker
12	solicitante		accem		réfugier		Migration
13	refúgio		apátrida		Exil		Immigration
14	refugiado		oua		subsidiaire		stateless
15	refugiar		refugiar		apatride		unaccompanied
16	reassentamento		immigration		traite		REFUGEES
17	latina		solicitantes		demandeur		Seekers
18	brasília		law		SPR		Resettlement
19	méxico		conare		OE		OFPRA

Fonte: elaboração própria.

Depois disso, iniciamos o processo (manual) de preenchimento das fichas. Cada palavra-chave deu origem a uma ficha-mãe, em que constam todas as suas UTEs, as quais foram coletadas por meio da ferramenta Word Sketches, que discutiremos a seguir.

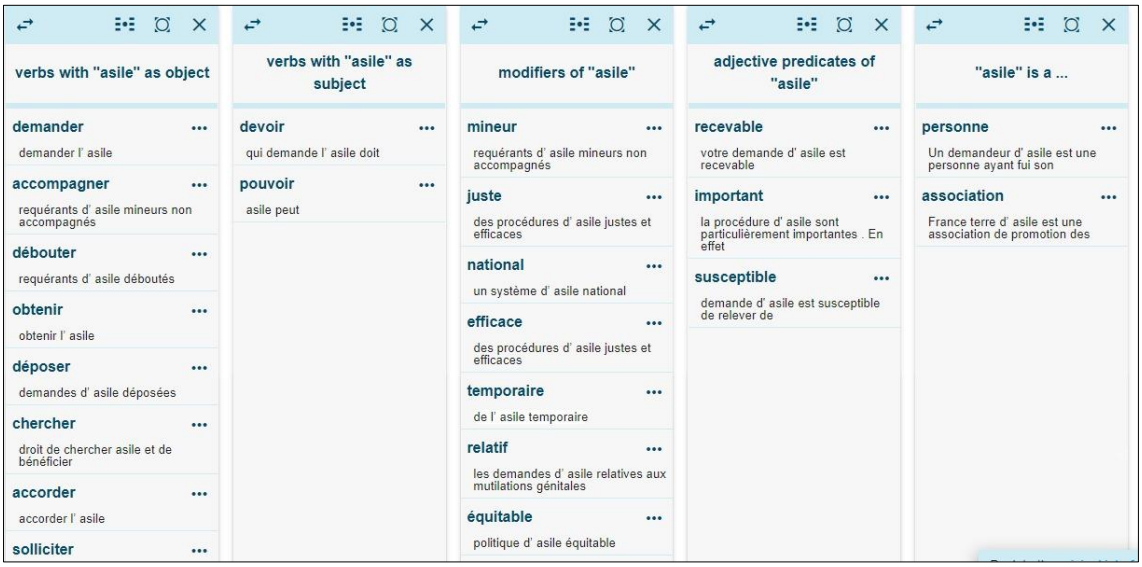
4.3.2. Coleta de UTEs com a ferramenta Word Sketches

A ferramenta Word Sketches[31] do SK processa os colocados e o contexto de ocorrência de uma palavra ou expressão de busca, podendo ser usada como uma espécie de sumário de seu comportamento colocacional e lexical. Os resultados são exibidos por

[31] SKETCH ENGINE. Word Sketch: Collocations and Word Combinations. Disponível em: https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/#toggle-id-4. Acesso em: 27 jun 2019.

categorias chamadas de *grammatical relations* (relações gramaticais). A Figura 5 mostra um exemplo do Word Sketch da palavra de busca “asile” no subcorpus de francês.

Figura 5 - Captura de tela do Word Sketch de "asile".



Fonte: elaboração própria.

A ferramenta ordena a lista de palavras colocando no topo as que possuem colocações mais típicas. Para determinar as “colocações típicas”, o programa utiliza a medida do LogDice[32], que compara as frequências de co-ocorrência de todas as palavras para calcular o *score* de colocação. Uma vez que os resultados são exibidos, é possível exportá-los em diversos formatos. No nosso caso, exportamos para o formato *.xls* do Excel, o qual exibe os dados da seguinte forma:

[32] SKETCH ENGINE. LogDice. Disponível em: https://www.sketchengine.eu/documentation/statistics-used-in-sketch-engine/#logdice. Acesso em: 27 jun 2019.

Figura 6 - Captura de tela do Word Sketches em que X equivale a "asylum aplication".

Grammar relation	Collocate	Freq	Score
modifiers of X	136	12.940	
	affirmative	11	8.160
	first	11	7.480
	time	6	7.080
	new	8	6.730
verbs with X as object	483	45.960	
	file	51	11.080
	lodge	42	10.970
	examine	46	10.850
	submit	28	9.970
	process	23	9.920
	make	57	9.380
	reject	16	9.320
	accept	12	9.090
	register	11	9.080
	receive	26	8.860
	approve	7	8.550
	adjudicate	6	8.540
verbs with X as subject	248	23.600	
	pend	7	9.210
	be	163	7.810
	have	41	7.490
	do	7	7.030
X and/or ...	106	10.090	
	decision	7	8.780
prepositional phrases	654	0.000	

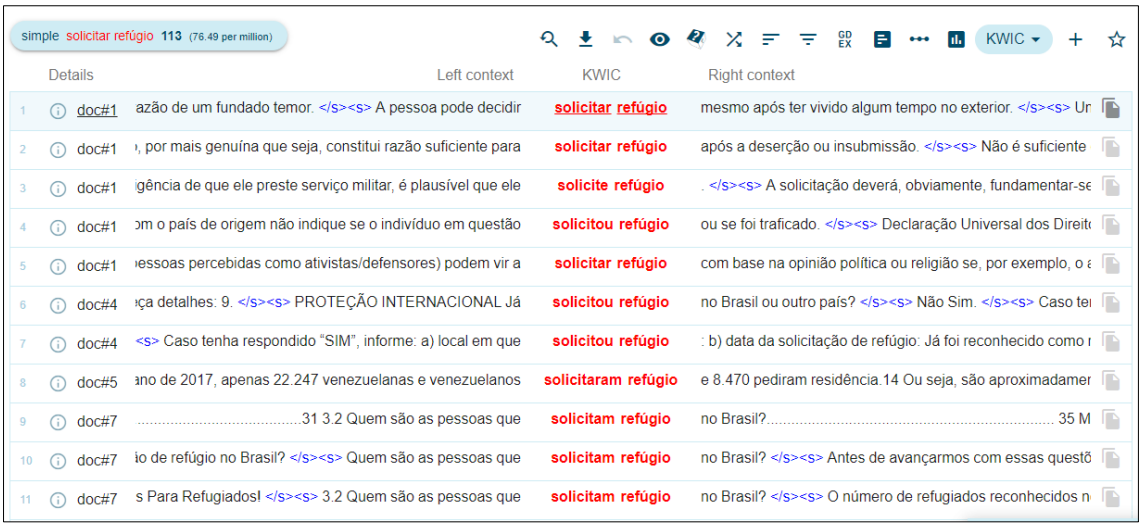
Fonte: elaboração própria.

Nessa página de exibição das combinações para a expressão de busca “asylum application”, Score corresponde à chavicidade, e Freq à frequência da expressão no corpus. Cada combinação listada, se verificada a ocorrência mínima de três vezes no corpus e em pelo menos três arquivos diferentes, gerou uma ficha-filha de UTE. Os dados provenientes destes arquivos do Word Sketches foram usados para preencher manualmente alguns campos da ficha, como por exemplo: a UTE (a combinação extraída, “**file an** asylum application” seria um exemplo) e a freq (frequência relativa da UTE no corpus). Os outros dados foram preenchidos com o auxílio da ferramenta Concordance.

4.3.3 Extração de dados com a ferramenta Concordance

Concordance[33] é uma ferramenta que exibe a palavra de busca nas linhas de contexto do corpus, e oferece diversas opções de consulta e ordenamento dos resultados. Abaixo, podemos ver um exemplo da expressão de busca "solicitar refúgio":

Figura 7 - Captura de tela da ferramenta Concordance.

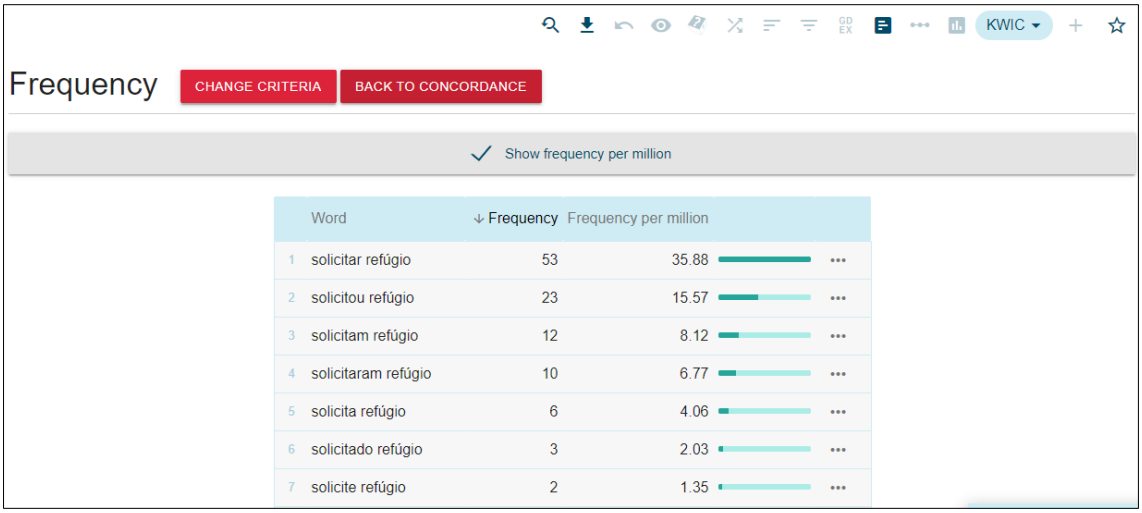


Fonte: elaboração própria.

A ferramenta oferece diversas opções avançadas, como a contagem de documentos em que a palavra de busca ocorreu e a seleção de termos que ocorrem até seis posições à esquerda ou à direita da expressão de busca, além de exibir, por exemplo, as variadas formas de ocorrência da expressão de busca lematizada (sem as flexões), como pode ser visto na Figura 8.

[33] SKETCH ENGINE. Concordance. Disponível em: https://www.sketchengine.eu/guide/concordance-the-most-powerful-tool-to-search-a-corpus/. Acesso em: 27 jun 2019.

Figura 8 - Captura de tela da opção de busca avançada para "solicitar refúgio" no Concordance.

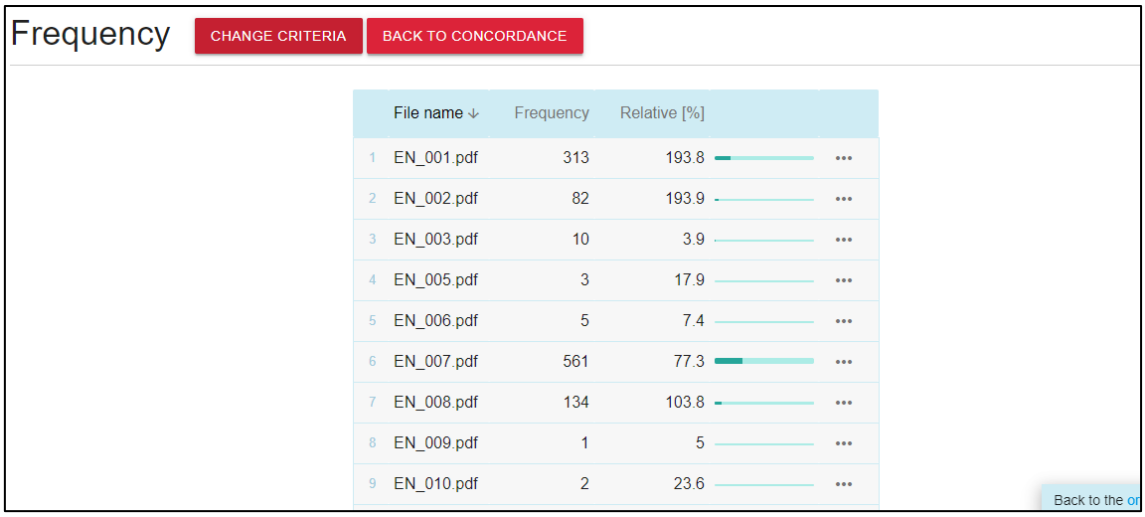


Fonte: elaboração própria.

As informações obtidas por meio do Concordance permitiram preencher manualmente os campos da ficha-filha para contexto de uso, observações e variantes. No campo UTEs relacionadas, colocamos UTEs que possuíam núcleo semelhante; por exemplo, na ficha da UTE “asylum application”, colocamos formas de combinação relacionadas: “**file an** asylum application”, “**lodge an** asylum application”, e assim por diante. Já nos campos de equivalência, utilizamos o mesmo método explicado no capítulo 4.3.1. para determinar a equivalência entre uma língua e outra.

Para o preenchimento das variantes, obtivemos uma lista dos arquivos em que a expressão de busca ocorreu e comparamos com uma tabela (disponível no Apêndice II) de controle em que listamos os nomes que os arquivos receberam, a quantidade de tokens e o país ao qual pertence. Todos os sub-corpus receberam essa rotina de extração de países no qual a variante ocorre, menos o sub-corpus do português, pois o único país representado é o Brasil. Abaixo, pode-se ver como o SK lista os arquivos em que uma busca ocorre.

Figura 9 - Lista de arquivos em que "asylum" ocorre no corpus.



Fonte: elaboração própria.

É importante frisar que, se uma UTE não tinha um equivalente *prima facie* ocorrendo de forma significativa no corpus (no mínimo três ocorrências em três arquivos diferentes, no nosso caso), ou se a busca pelo colocado não resultou em equivalentes possíveis com frequência maior que três, não foi possível determinar uma tradução, pois nosso estudo parte de uma abordagem dirigida por corpus. Uma solução seria ampliar o corpus para possibilitar maior amostragem da UTE em questão, ou acrescentar mais textos paralelos (original e tradução alinhadas no nível da sentença) confiáveis.

Passaremos agora para a uma explicação de como estruturamos as fichas no banco de dados criado.

5. ESTRUTURAÇÃO DO BANCO DE DADOS COM YAML E PYTHON

Neste capítulo, vamos expor como foi estruturado o banco de dados e como o programa de busca de UTEs foi programado. Primeiramente, apresentamos a linguagem de programação e o formato de arquivos utilizados – Python e YAML. Depois, descrevemos os requisitos necessários para a construção do banco de dados e, por fim, explanamos a lógica de estruturação deste.

5.1. YAML e PYTHON

O YAML[34], acrônimo para *YAML Ain't Markup Language*[35], é um formato de codificação de dados estruturados que garante ótima legibilidade para humanos. Foi criado essencialmente para armazenamento de dados, possui boa integrabilidade com linguagens de programação, tem um modelo consistente de formatação e é de fácil usabilidade. O YAML foi escolhido graças a sua habilidade de codificar os dados em utf-8, além de ser facilmente editável por humanos. Além disso, é possível integra-lo à linguagem de programação que escolhemos, o Python. Sendo assim, determinamos que as nossas fichas deveriam ser codificadas em utf-8 por meio do YAML.

O Python[36], por sua vez, é uma linguagem de programação orientada a objetos popular e muito flexível, ideal para ser usada em análises linguísticas. É de uso gratuito, com dados abertos e gerenciada pela ONG Python Software Foundation[37]. Escolhemos o Python pela facilidade de manipulação, além de ser a linguagem de programação ensinada na disciplina de Língua e Programação, ministrada na graduação de LEA-MSI.

[34] YAML. Disponível em: https://yaml.org/. Acesso em: 28 jun 2019.
[35] "YAML não é uma linguagem de marcação". Tradução nossa.
[36] PYTHON. Disponível em: https://www.python.org/. Acesso em: 28 jun 2019.
[37] PYTHON SOFTWARE FOUNDATION. About. Disponível em : https://www.python.org/psf/. Acesso em: 28 jun 2019.

Por ser somente um formato de codificação, o YAML não está inserido dentro do Python. Para fazer com que ambos trabalhassem juntos, precisamos instalar uma biblioteca no Python chamada Pyyaml[38]. Essa biblioteca faz a ponte entre o Python e o YAML, permitindo construir um programa em Python que leia arquivos em YAML, por exemplo.

Para a construção de qualquer produto que tenha como usuário final uma pessoa, é necessário o uso de boas práticas, como por exemplo, o levantamento de requisitos. Nesse sentido, na próxima seção descrevemos os requisitos para a construção do banco de dados utilizado neste trabalho, representando-o com os fluxogramas que guiaram o processo de programação e tomada de decisões.

5.1.2. Levantamento de requisitos

O levantamento de requisitos é uma etapa essencial na construção de qualquer projeto de programação. Consiste na determinação de funcionalidades e tarefas que o aplicativo deve desempenhar até que possa chegar ao usuário final.

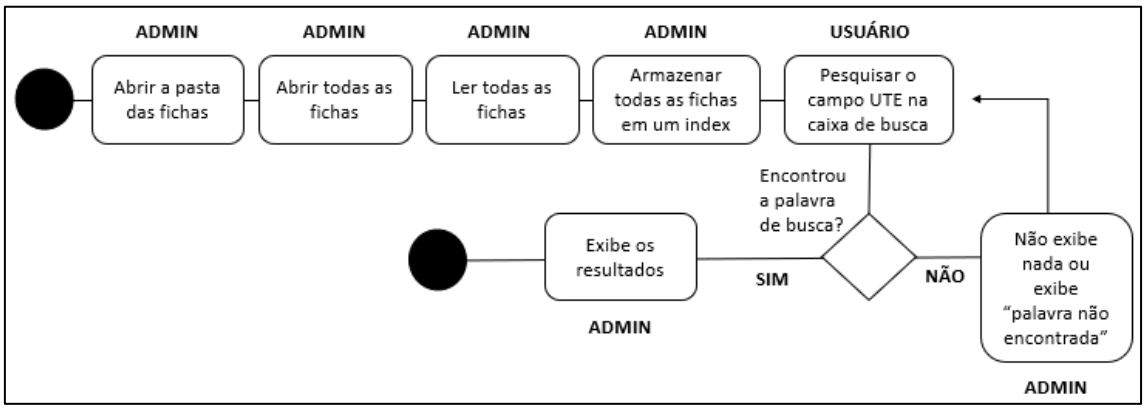
O primeiro parâmetro que tivemos que decidir foi em qual versão do Python deveríamos programar nosso sistema. Atualmente, o Python está na versão 3.7.3, mas escolhemos trabalhar com a versão 2.7, que é mais estável, pois temos como objetivo final entregar um aplicativo que possa ser acessado em uma plataforma online. Além disso, escolhemos esta versão pois os recursos que utilizamos (explicados a seguir) não fornecem suporte para a versão do Python. A plataforma que escolhemos para hospedagem foi o App Engine[39] da Google, uma plataforma de criação e hospedagem de

[38] PYYAML. Disponível em: https://pypi.org/project/PyYAML/. Acessado em: 28 jun 2019.
[39] GOOGLE APP ENGINE. Disponível em: https://cloud.google.com/appengine/?hl=pt-BR. Acessado em 28 jun 2019.

aplicativos em que não é necessário possuir servidor. O Google oferece uma quota de serviço gratuito, no qual é possível hospedar um aplicativo como o nosso e mantê-lo online de forma gratuita, e foi esse o fator determinante da nossa escolha.

Para a construção do aplicativo, vamos simular como as tarefas deveriam ser desempenhadas e em que ordem deveriam ser feitas. O fluxograma da Figura 10 representa as tarefas que devem ser executadas internamente pelo aplicativo.

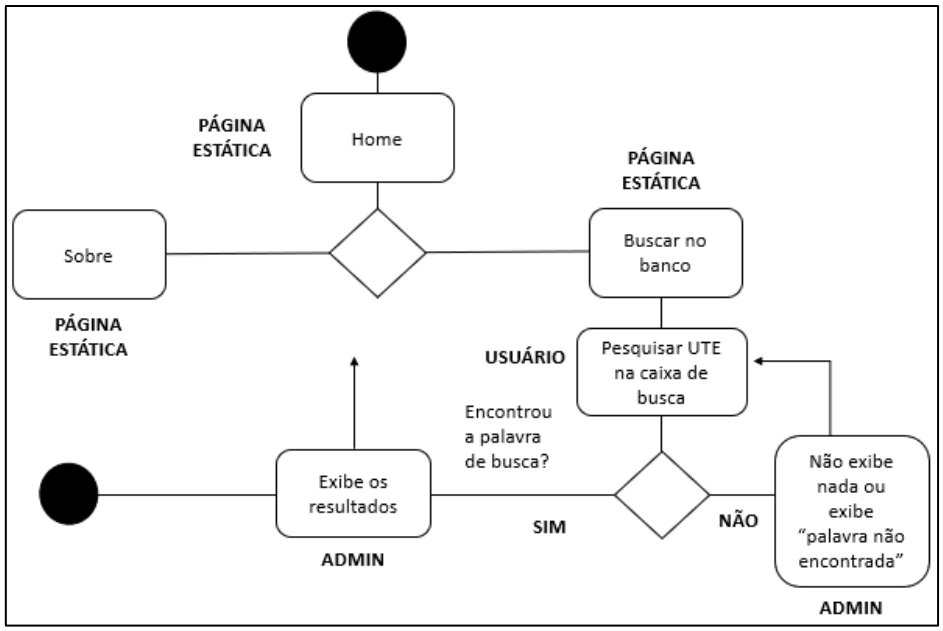
Figura 10 - Fluxograma das atividades que devem ser desempenhadas internamente pelo aplicativo.



Fonte: elaboração própria.

Uma vez concluído o fluxograma das tarefas internas, precisamos pensar em como o usuário iria pesquisar as entradas do nosso glossário e como isso poderia ser feito em uma página online. Criamos um segundo fluxograma, que pode ser visto na Figura 11, para descrever a jornada do usuário no site.

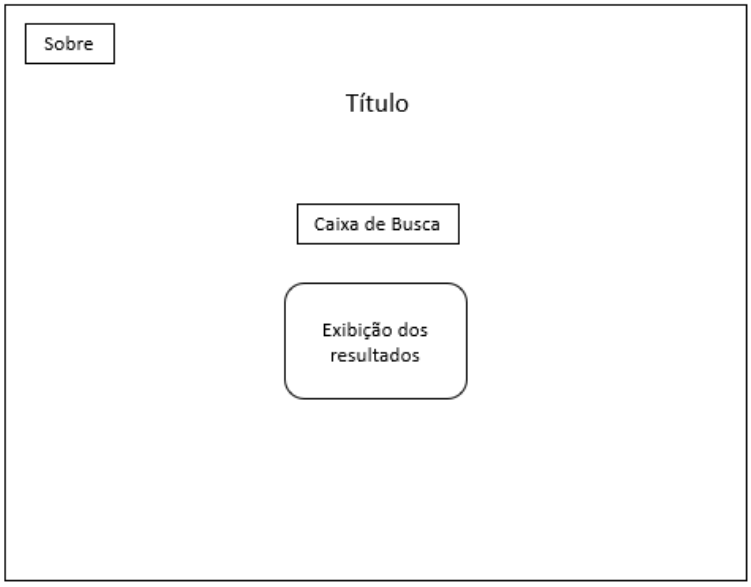
Figura 11 - Fluxograma da jornada do usuário no site.



Fonte: elaboração própria.

Em seguida, elaboramos um esboço do site que pudesse guiar a forma como faríamos a disposição dos elementos para que o usuário possa acessar ao conteúdo das fichas posteriormente.

Figura 12 - Esboço da interface do site.



Fonte: elaboração própria.

Após a montagem do esboço, nos deparamos com a necessidade de fazer uma ponte entre o Python e o site de busca. Para isso, utilizamos o Jinja, e para a composição do site, o HTML 5, sobre os quais falaremos na próxima seção.

5.2 HTML 5 e Jinja

O *HyperText Markup Language*[40] – HTML é uma linguagem de marcação para construir sites na web. Os arquivos de HTML são interpretados pelos navegadores que utilizamos para acessar a internet. Atualmente, a linguagem se encontra na versão 5 e possibilita distintas opções de visualização gráfica, como o uso de efeitos 3-D[41]. O HTML pode ser facilmente combinado com algumas linguagens de programação, como o Java, o Ruby, o PHP e, no nosso caso, o Python. Para fazer a conexão entre o HTML e esta linguagem de programação, utilizamos o Jinja.

O Jinja[42] é um mecanismo de templates para o Python. É ele quem permite que o HTML reconheça o código em Python no arquivo de marcação HTML. Além disso, permite que seja possível chamar objetos do Python no código HTML, o que é de extrema importância, pois o Python é uma linguagem orientada a objetos. Para utilizá-lo, basta instalar sua biblioteca no Python e importar a função quando for utilizar. Aqui é importante lembrar que o Jinja não tem suporte para o Python 3, e esse é mais um dos motivos pelos quais estamos usando o Python 2.7.

Finalmente, o esquema de cores e todo o design do site foi definido em uma folha de estilos *Cascading Style Sheets* – CSS. O HTML não precisa de muito para reconhecer a folha de estilos, basta adicionar uma referência a ela no código da página em HTML.

---

[40] Linguagem de Marcação de Hipertexto.
[41] W3SCHOOLS. HTML 5 Introduction. Disponível em: https://www.w3schools.com/html/html5_intro.asp, Acessado em: 28 jun 2019.
[42] JINJA 2. Docs. Disponível em : http://jinja.pocoo.org/docs/2.10/. Acesso em: 28 jun 2019.

5.3 Usando o indexador do Google Cloud

Para buscar as informações contidas nas fichas, precisamos indexa-las utilizando um indexador, que pode ser criado com programação ou importado já pronto, como no nosso caso.

Como optamos por utilizar os serviços do App Engine, o Google disponibiliza o seu próprio indexador[43] para os aplicativos que lá se hospedam. Para utiliza-lo, basta importar a função "search" da biblioteca do Google Cloud e utilizar seus atributos para montar o aplicativo que se deseja fazer.

O API Search, nome dado ao indexador, é indicado para armazenar documentos com dados estruturados, justamente o nosso caso com as fichas em YAML. Não há limite para a indexação de documentos; entretanto, o Google oferece uma quota gratuita de serviços que pode restringir a quantidade de dados por tamanho. É crucial destacar que, até o momento em que esse trabalho foi finalizado, o API Search só estava disponível para o Python 2.7, o que nos deu mais um motivo para utilizar essa versão da linguagem de programação.

A seguir, apresentamos a macro e a microestrutura do nosso glossário digital.

[43] GOOGLE CLOUD. Documentos e Índices. Disponível em: https://cloud.google.com/appengine/docs/standard/python/search/?hl=pt-br . Acesso em: 28 jun 2019.

6. MACRO E MICRO ESTRUTURA DO GLOSSÁRIO DIGITAL

No capítulo anterior, apresentamos um esboço de site para implementação do programa de busca que criamos. Como o nosso glossário é um repertório especializado em formato digital, determinamos o que cada verbete deveria conter e como o usuário tem acesso a eles no site de consulta. Como o nosso projeto foi desenvolvido em muito pouco tempo, sugeriremos um modelo de funções mínimas, que poderão ser modificadas posteriormente.

6.1. A macroestrutura do glossário

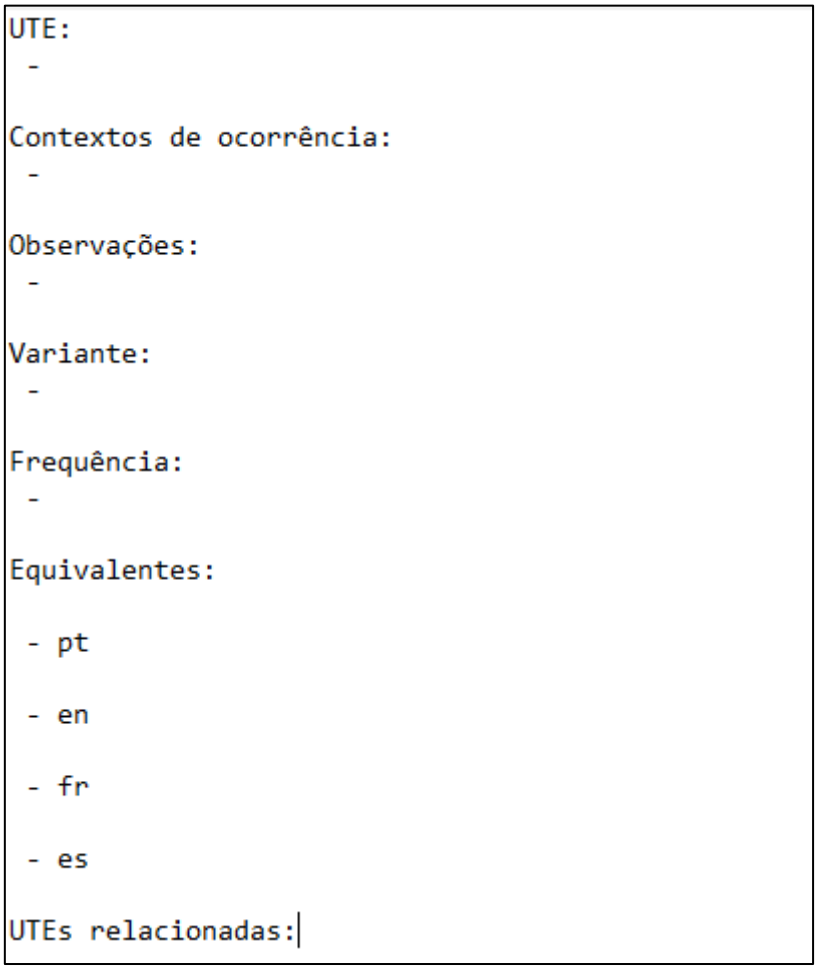
Como mostrado no capítulo 3, a macroestrutura consiste na forma como um glossário é apresentado ao público-alvo. Para compor a macroestrutura do nosso glossário, elencamos as seguintes estruturas:

a) **Página "Sobre"** – Deve exibir informações sobre o projeto, como o nosso glossário foi feito, como ele funciona e a equipe de trabalho;
b) **Página "Inicial"** – Exibe a caixa de busca por meio da qual os usuários podem ter acesso aos verbetes;
c) **Rodapé com "Contato"** – Colocado no fim da página, fornece os dados de contato da equipe, caso o usuário queira solicitar mais informações.
d) **Organização dos verbetes** – Os verbetes pesquisados aparecerão em ordem alfabética se contidos no banco de dados.

6.2. A microestrutura dos verbetes

Assim como o site tem uma macroestrutura, também deve seguir uma microestrutura, que é como as informações da ficha serão exibidas. A seguir, mostramos o modelo de exibição do nosso verbete:

Figura 13 - Campos presentes na microestrutura do nosso glossário.



Fonte: elaboração própria.

Aqui, é importante explicar como se dá o método de exibição. O programa tem uma função automática – "results" – herdada da função "search", cujo produto final exibe os resultados de forma aleatória, ou seja, desconfigurando a estrutura desenhada nas fichas. Por exemplo, quando busco por "refúgio", de acordo com o modelo, a informação que deveria ser exibida em seguida é o contexto de uso. Entretanto, isso não acontece de forma automática, pois a função ordena seus resultados de forma aleatória. Por esse motivo, montamos a estrutura acima e destacamos que optamos por ocultar alguns

campos presentes nas fichas, como as definições (que não utilizamos) e o id (que é só para controle interno).

Ainda assim, configuramos o programa de sorte a não exibir os campos **obs**, **pt**, **en**, **fr** e/ou **es** no caso de estes campos não conterem nenhuma informação, deixando a visualização mais limpa e organizada. Ainda que seja recomendado manter a consistência na microestrutura das fichas, apresentando sempre os mesmos dados, optamos por isso pois não podemos sugerir traduções para uma UTE que não tenham sido identificadas no corpus seguindo os princípios e a metodologia em LC utilizada.

A seguir, apresentamos informações gerais sobre o corpus compilado e os dados obtidos com a metodologia proposta para, por fim, analisar os resultados do glossário e tecer as considerações finais.

7. RESULTADOS

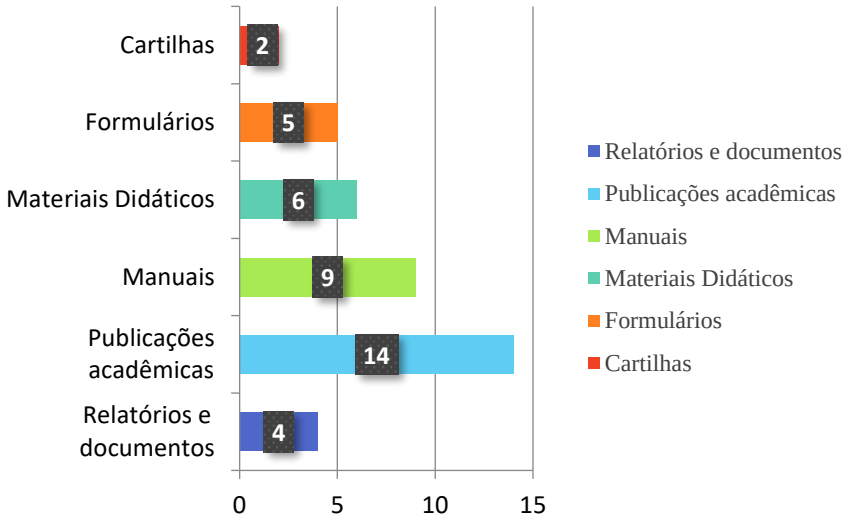
7.1. Dados sobre o COMMIRE

A seguir, mostramos informações detalhadas sobre a composição dos sub-corpus coletados para compor o COMMIRE – **Co**rpus **M**ultilíngue de **Mi**gração e **Re**fúgio, dando exemplos de resultados extraídos com ferramentas utilizadas e fichas preenchidas. Adiante, mostraremos como é a visualização do glossário digital e como funciona o mecanismo de busca com capturas de tela.

7.1.1. Tipologia textual do subcorpus do português

Como exposto na Tabela 1, o objetivo inicial de coleta do corpus era recolher materiais aos quais os refugiados, os solicitantes de refúgio e os imigrantes têm acesso quando chegam ao país onde vão solicitar refúgio ou para o qual vão, eventualmente, imigrar. Geralmente, esses materiais são cartilhas e revistas informativas, formulários que devem ser preenchidos ou manuais. Nesse sentido, apresento na figura abaixo o gráfico de composição textual do subcorpus de português brasileiro (por quantidade de arquivos), primeiramente.

Gráfico 1 - Composição do corpus em português por quantidade de arquivos.

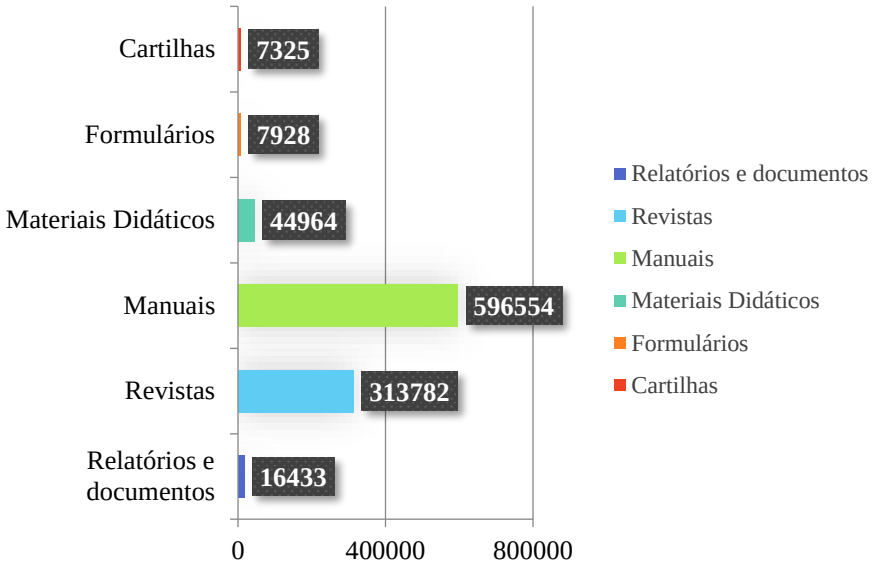


Fonte: elaboração própria.

De acordo com o Gráfico 1, é possível notar que a maior quantidade de material do corpus em português, composto de 36 arquivos, é de publicações acadêmicas. Isso pode ser, em primeiro lugar, porque o Brasil produziu poucas cartilhas para os refugiados e solicitantes de refúgio (somente duas), mas há uma quantidade considerável de manuais e materiais para os agentes que trabalham diariamente com os refugiados e solicitantes.

Abaixo, apresentamos um gráfico da composição do corpus do português em relação à quantidade de palavras. Novamente, os manuais e as publicações acadêmicas, as quais consistem em cadernos de ensaios produzidos por diversos especialistas da área, são os materiais que apresentam o maior número de palavras, por serem materiais que apresentam grande quantidade de informação e textos maiores.

Gráfico 2 - Composição do corpus por quantidade de palavras.



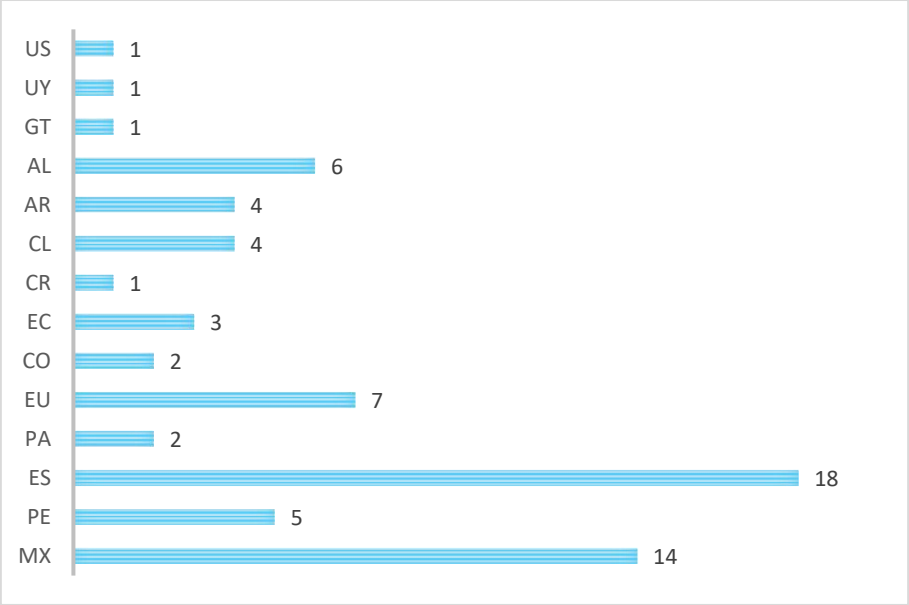
Fonte: elaboração própria

7.1.2. Tipologia textual do subcorpus do espanhol

O subcorpus do espanhol é composto de 70 arquivos de diferentes países hispano falantes. A saber: México, Peru, Espanha, Colômbia, Equador, Costa Rica, Chile, Argentina, Panamá, Guatemala e Uruguai. Há ainda documentos que foram produzidos por Organizações Internacionais (OIs), e por isso não possuem nacionalidade específica e foram etiquetados como pertencentes à União Europeia ou à América Latina, no caso de terem sido produzidos em algum desses lugares. Essa decisão foi tomada com o objetivo de obter maior representatividade linguística no corpus, além de abarcar um maior número de variações, fazendo com que os resultados sejam mais confiáveis. Na figura abaixo, é possível observar a quantidade de arquivos por países, representados por suas siglas padronizadas pelo ISO 3166-1[44] (a lista de códigos utilizados nesta monografia encontra-se na Apêndice 1 deste trabalho).

[44] Organisation Internationale de Normalisation (ISO). **Country Codes**. Disponível em: https://www.iso.org/obp/ui/#home. Acesso em: 20 jun 2019.

Gráfico 3 - Quantidade de arquivos no corpus do espanhol divididos por país.



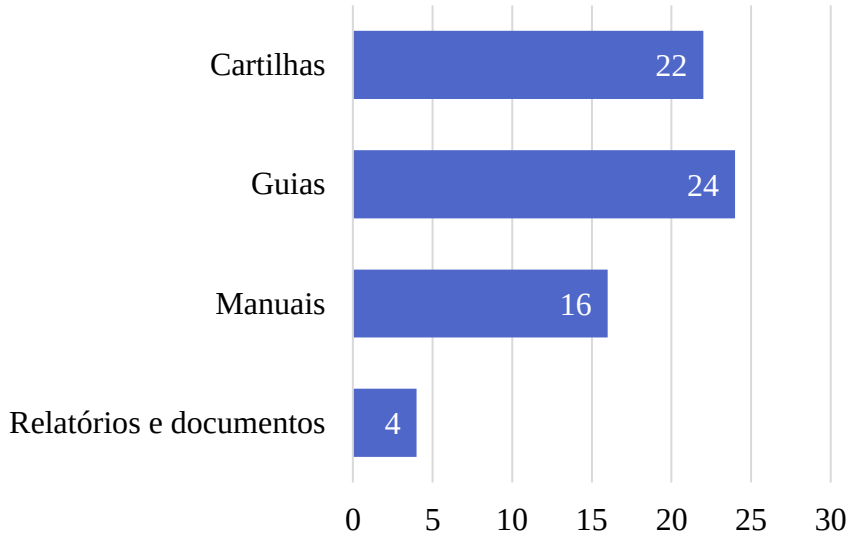
Fonte: elaboração própria.

Na figura acima, é importante destacar a presença de um arquivo produzido nos Estados Unidos, que é uma tradução de um arquivo em inglês. Foi incluído nos arquivos pois foi encontrado entre as primeiras posições de busca do Google, o que significa que deve ser baixado pelo público de interesse com frequência.

Uma questão a ser comentada é a presença desses arquivos na internet. Os governos dos países não deixam claro até que ponto e o quanto investem na produção de materiais para o acolhimento de refugiados e imigrantes, mas é possível imaginar que países que recebem maiores quantidades de solicitação, investirão mais na produção e disseminação desses materiais na internet, ainda que muitos deles sejam produzidos por ONGs ou por OIs. Dessa maneira, é possível notar que o México (MX) e a Espanha são os países que mais possuem representatividade no corpus.

Na figura a seguir, apresentamos um gráfico com a composição da tipologia textual do subcorpus do espanhol:

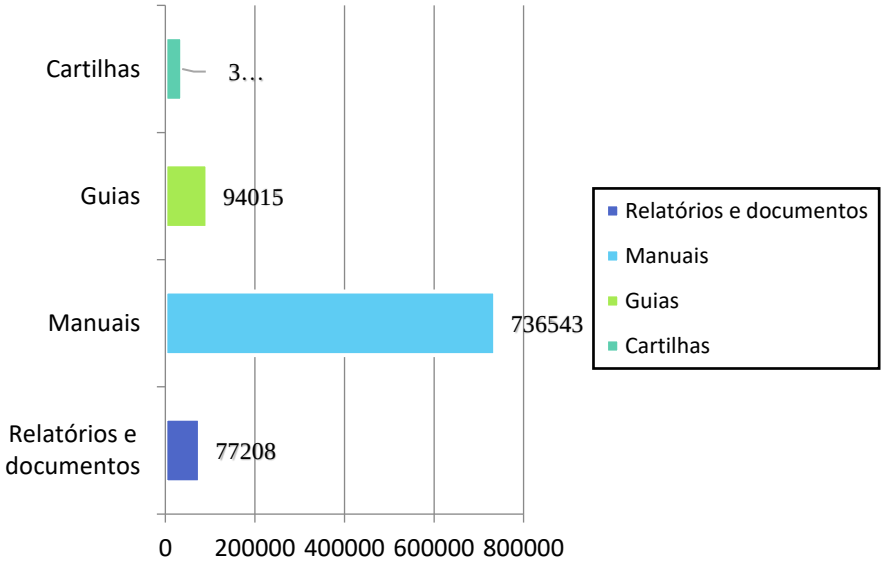
Gráfico 4 - Quantidade de arquivos por tipologia textual do corpus do espanhol.



Fonte: elaboração própria.

Como o corpus do espanhol é composto de múltiplas variedades do espanhol, foi muito mais fácil encontrar cartilhas e guias do que no português, por exemplo. Por isso, não foi necessário incluir revistas de divulgação científica nem formulários no corpus. Ainda assim, como dito anteriormente, um dos critério de seleção de textos que utilizamos diz respeito à facilidade de se encontrar o material, pois para produzir um glossário que vai ser efetivamente utilizado, é importante que os tradutores, interpretes e até mesmo os imigrantes acessem a terminologia presente nos materiais que vão consultar/ traduzir. No Gráfico 4 pode-se observar a quantidade de palavras por tipologia textual no subcorpus do espanhol.

Gráfico 5 - Quantidade de palavras por tipologia textual no corpus do espanhol.

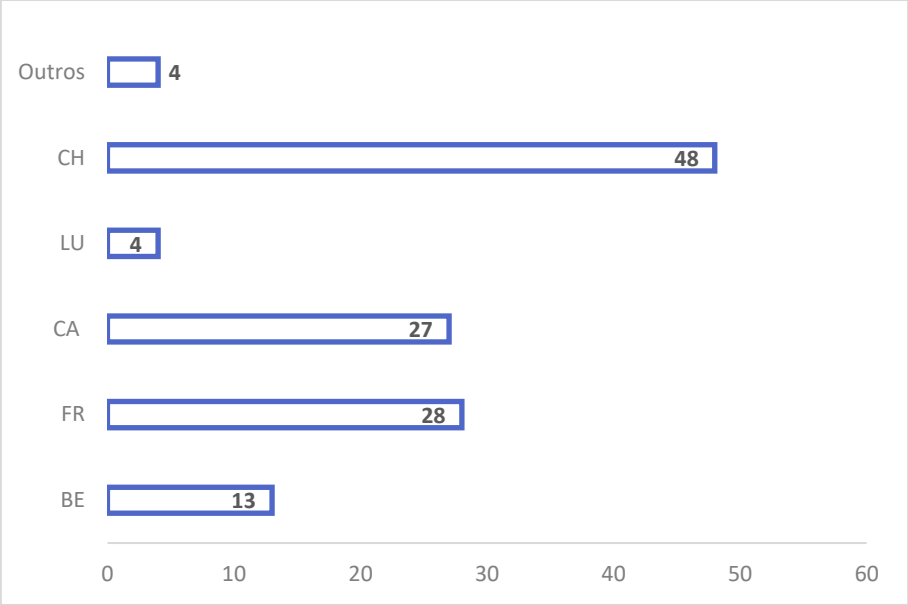


Fonte: elaboração própria.

7.1.3 Tipologia textual do subcorpus do francês

O corpus do francês é constituído de 131 arquivos de uma variedade de países francófonos. São eles: Bélgica, França, Canadá, Luxemburgo e Suíça. É importante mencionar que foram procurados arquivos relacionados à África francófona. Alguns foram encontrados, porém não em francês, mas em árabe e outras línguas minoritárias, o que impossibilitou a inclusão no corpus. Abaixo, é possível ver a distribuição de arquivos por país (também representados pelas siglas ISO).

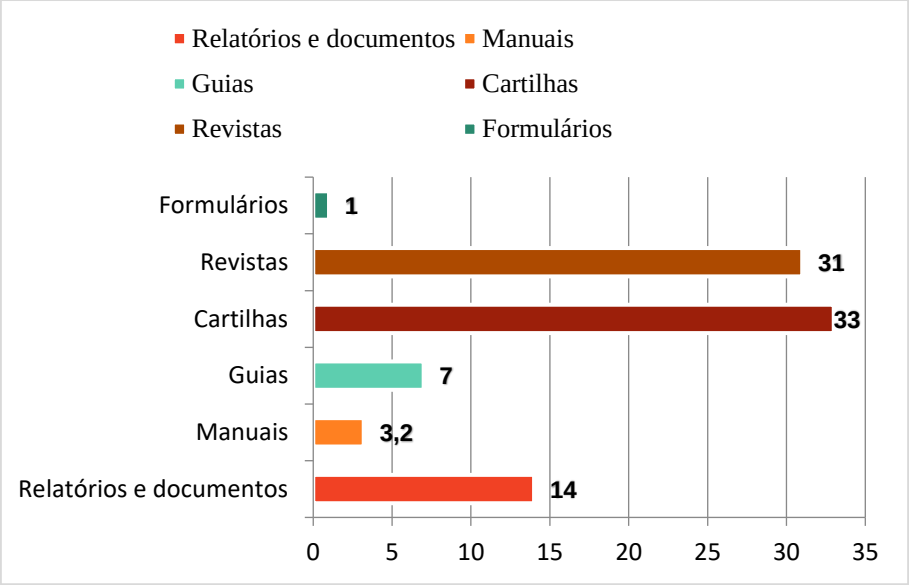
Gráfico 6 - Arquivos distribuídos por país no corpus da língua francesa.



Fonte: elaboração própria.

Há uma grande quantidade de arquivos da Suíça (CH), o que se deve principalmente ao fato de que a sede da ONU é localizada na capital do país, Genebra, e provavelmente há uma maior produção de materiais ali. Um outro ponto que merece destaque é a etiqueta “outros”, que abarca a quantidade de arquivos declaradamente traduzidos para o francês. Um deles, por exemplo, é um documento em francês coletado em um site português (PT). No arquivo, a ONG que o publicou disse que foi traduzido por voluntários afro-francófonos. Os demais arquivos são traduções de documentos produzidos no âmbito da União Europeia.

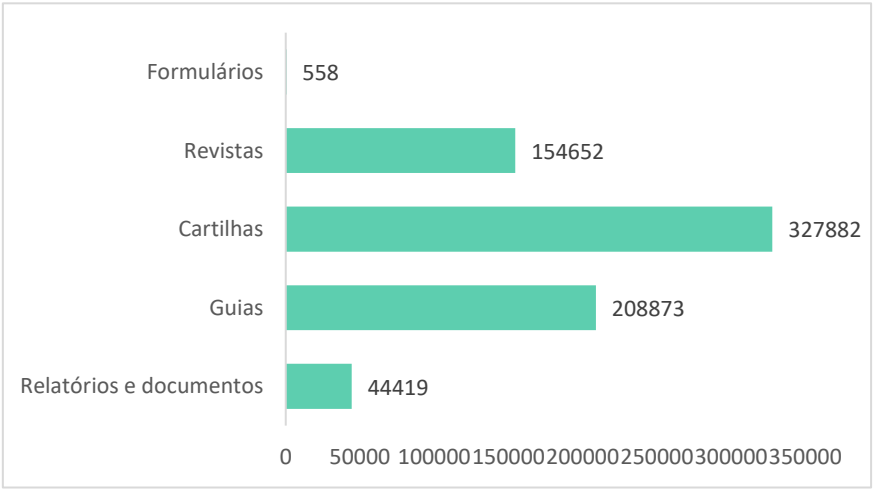
Gráfico 7 - Tipologia textual dos arquivos do corpus francês.



Fonte: elaboração própria

Há uma grande quantidade de revistas no corpus, tais como a *Planète Éxil*, produzida em vários volumes em parceria às OIs, em especial a ONU. A revista convida especialistas da área (de diversas nacionalidades), refugiados e solicitantes de refúgio para falar sobre o tema, e tem como editores voluntários francófonos. No gráfico a seguir, mostramos a quantidade de palavras por tipologia textual.

Gráfico 8 - Quantidade de palavras por tipologia textual no corpus francês.



Fonte: elaboração própria.

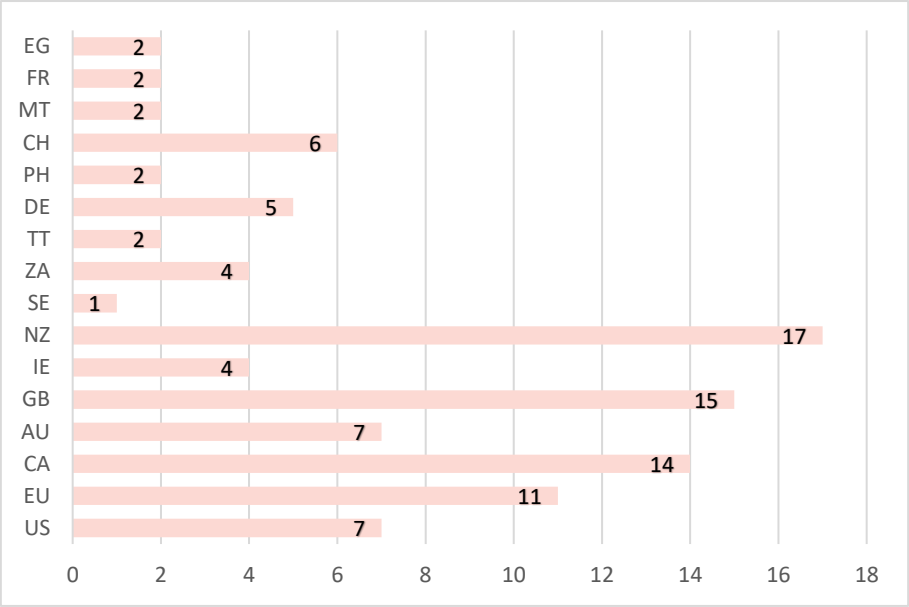
Aqui é importante pontuar que não há muita distinção entre os guias e as cartilhas, visto que ambos visam informar alguém sobre algum assunto. Entretanto, classificamos como guia todo material que oferecia informações sem aspectos visuais muito elaborados, ao passo que a cartilha constitui um documento mais visual. Ainda sobre isso, há documentos no corpus que poderiam ser facilmente classificados como guias, entretanto não o foram por não terem sido apresentados em documentos formatados (*.pdf* ou *e-book*), mas sim como uma postagem ou texto corrido em prosa em uma página. Embora os arquivos tenham maneiras divergentes de se apresentar ao leitor, todos tem o objetivo de informar, exceto pelos formulários.

7.1.4 Tipologia textual do subcorpus em inglês

O corpus do inglês é composto de 101 arquivos coletados entre os seguintes países: Estados Unidos, Canadá, Austrália, Reino Unido (Inglaterra, Irlanda do Norte, Grã-Bretanha e Escócia), Irlanda, Nova Zelândia, Suécia, África do Sul, Trinidade e Tobago, Alemanha, Filipinas, Suíça, Malta e Egito.

Aqui, devemos nos lembrar que o inglês é visto atualmente como língua franca (HOFFMAN, 2000). Por causa disso, a presença do inglês na internet é muito maior do que as outras línguas, e, por isso, há mais conteúdo sendo produzido e ofertado nesse idioma. A seguir, mostramos a quantidade de arquivos por país, representado pelos seus respectivos códigos ISO.

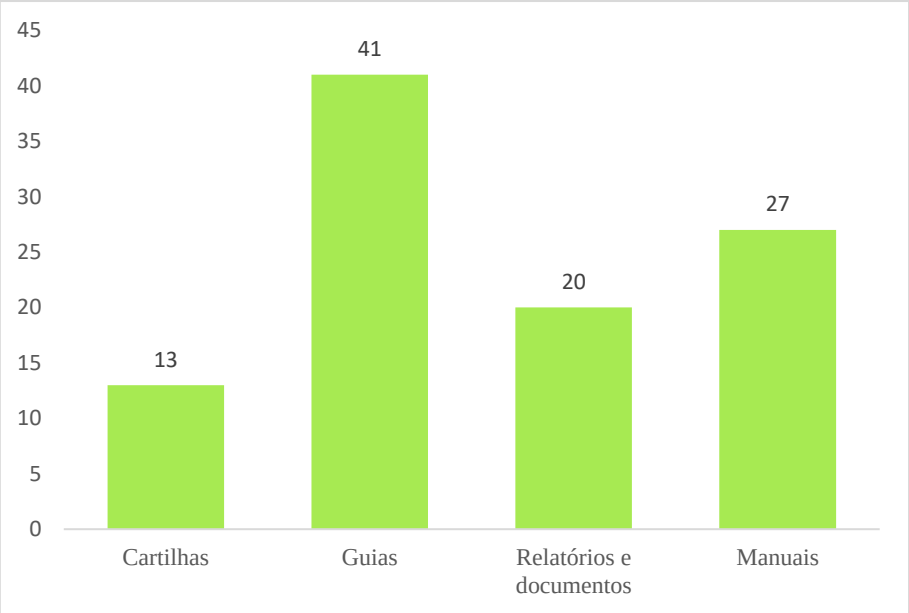
Gráfico 9 - Quantidade de arquivos por país no corpus inglês.



Fonte: elaboração própria.

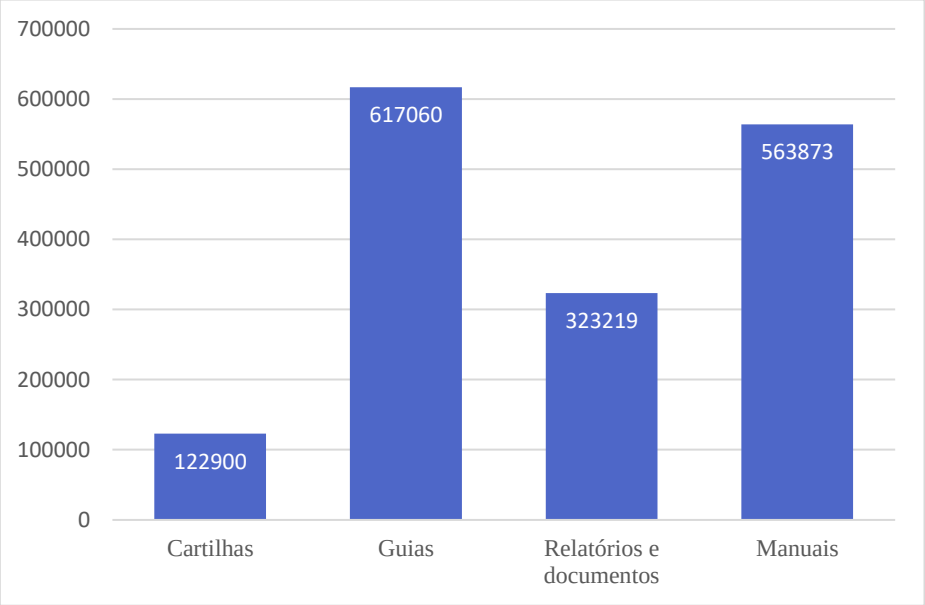
Como se pode observar acima, há dois arquivos em inglês, coletados em sites franceses. Estes documentos são de traduções de duas cartilhas que aparecem no topo das buscas quando se procura por "cartilhas para solicitantes de refúgio". Ainda assim, verificamos a presença da sigla EU, que se refere à *European Union* – União Europeia. A estes arquivos foi dada esta sigla pelo fato de terem sido produzidos no âmbito do bloco internacional e discorrerem sobre o processo de refúgio no bloco como um todo. Vejamos agora a quantidade de arquivos por tipologia textual, e a quantidade de palavras por tipologia textual, nas figuras a seguir, respectivamente.

Gráfico 10 - Quantidade de arquivos por tipologia textual no corpus do inglês.



Fonte: elaboração própria.

Gráfico 11 - Quantidade de palavras por gênero textual.



Fonte: elaboração própria.

Podemos verificar que o subcorpus de inglês é composto em sua maioria por guias e manuais. O inglês, por seu status de língua franca, tem uma maior oferta de materiais, tanto para o imigrante, quanto para quem trabalha com ele.

Por fim, tivemos a seguinte quantidade total de palavras-forma (*types*) e palavras-ocorrência (*tokens*):

Tabela 3 - Quantidade de tokens e types por corpus.

Língua do Corpus	**Tokens**	**Types**
Português	1,477,349	54,583
Espanhol	1,305,011	47,437
Francês	1,306,553	40,866
Inglês	2,091,482	57,332

Fonte: elaboração própria.

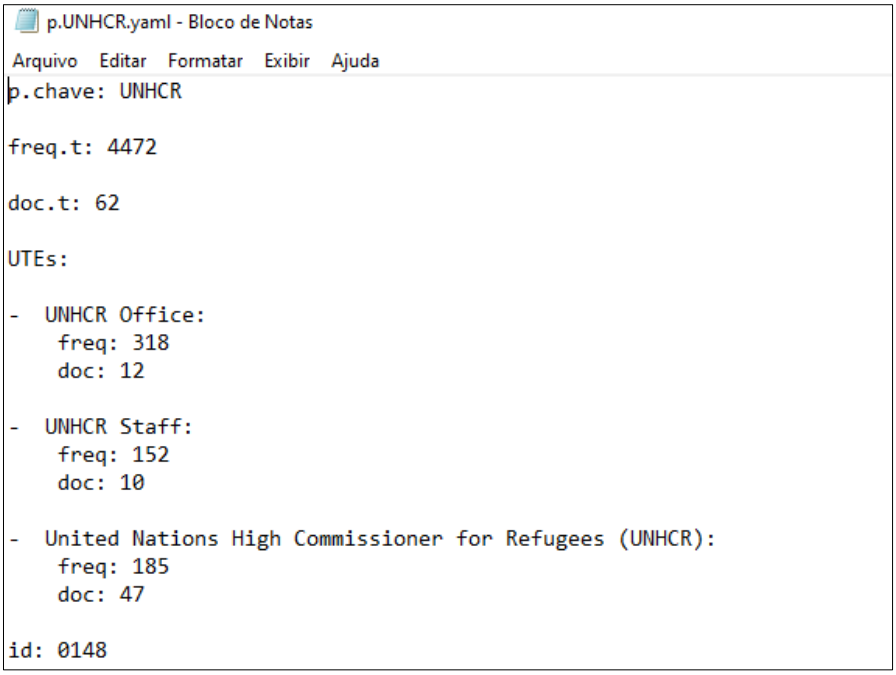
É importante mencionar há distinção entre a quantidade de tokens e palavras pois a maioria dos arquivos foram inseridos no SK como *.pdf,* o programa utiliza os metadados dos arquivos para calcular o número de palavras do corpus e mede a quantidade de tokens pela quantidade de espaços entre as palavras. No processo de reconhecimento óptico, muitas incongruências ocorreram, o que acarretou ruído no corpus e por isso há uma maior contagem de tokens do que de palavras.

Passemos agora para as ferramentas utilizadas para a extração das UTEs.

7.2 Exemplos de fichas de UTEs preenchidas

Depois de coletados todos os dados do SK, preenchemos manualmente cada ficha. As Figuras 24 e 25 mostram, respectivamente, um exemplo de ficha-mãe da palavra-chave UNHCR e de ficha-filha (de UTE) preenchidas:

Figura 14 - Captura de tela da ficha da palavra-chave “UNHCR”.



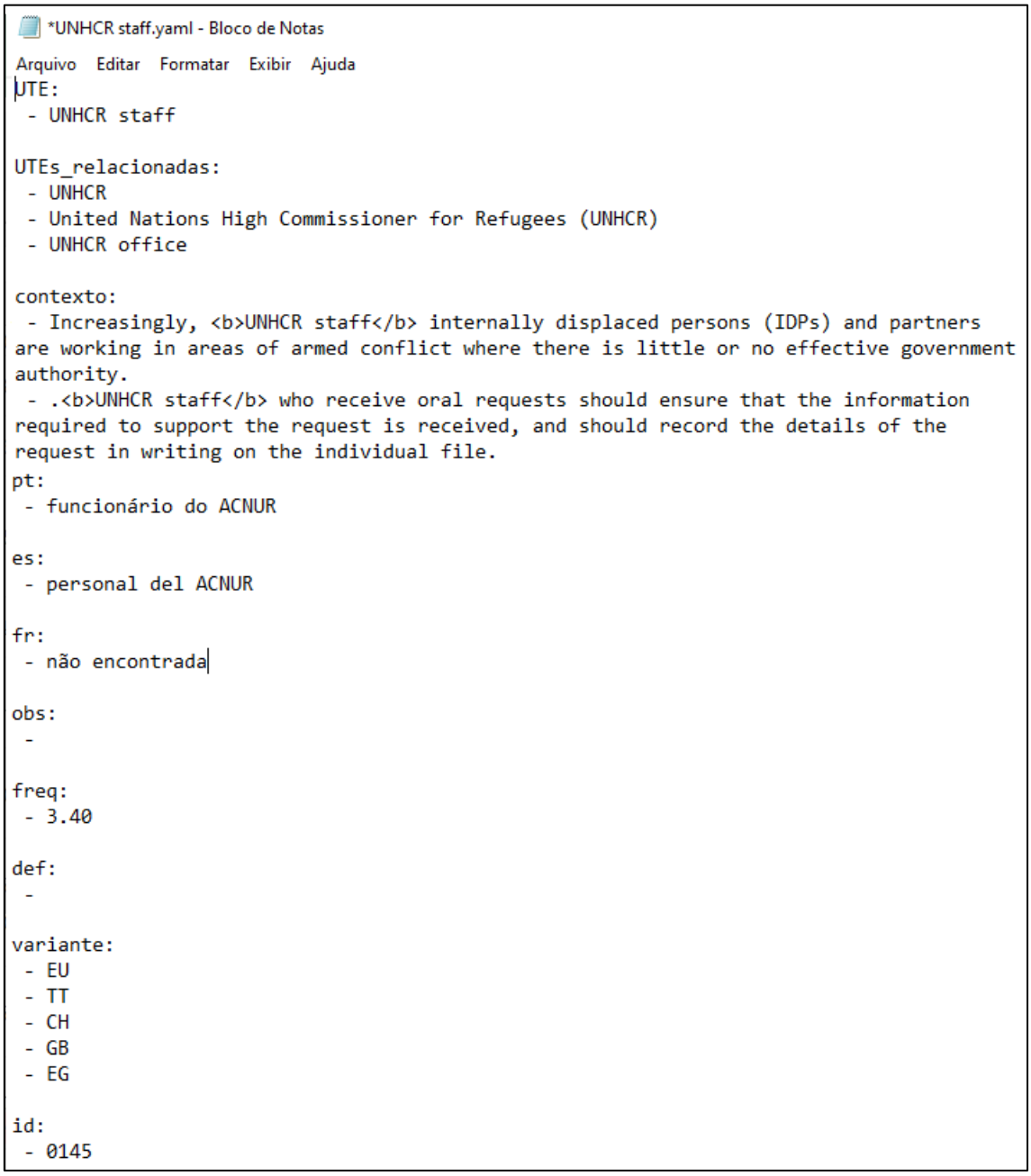
Fonte: elaboração própria

Figura 15 - Captura de tela da ficha da UTE "UNHCR office"

```
UTE:
 - UNHCR Office

UTEs_relacionadas:
 - UNHCR
 - asylum staff
 - United Nations High Commissioner for Refugees (UNHCR)
 - asylum office

contexto:
 - The specific procedures adopted by each <b>UNHCR Office</b> will necessarily
reflect the size of the particular RSD operation, the staffing and other resources
available in the UNHCR Office, as well as the conditions in the particular country.
 - In each <b>UNHCR Office</b>, a Protection staff member should be designated to act
as the Protection focal point for security issues in the Office.

pt:
 - escritório do ACNUR

es:
 - oficina del ACNUR
fr:
 - não encontrada

obs:
 -

freq:
 - 7.11

def:
 -

variante:
 - CA
 - EU
 - AU
 - NZ
 - CH
 - MT
 - EG
 - TT
 - GB

id:
 - 0144
```

Fonte: elaboração própria

Figura 16 - Captura de tela da ficha de UTE "UNHCR staff."



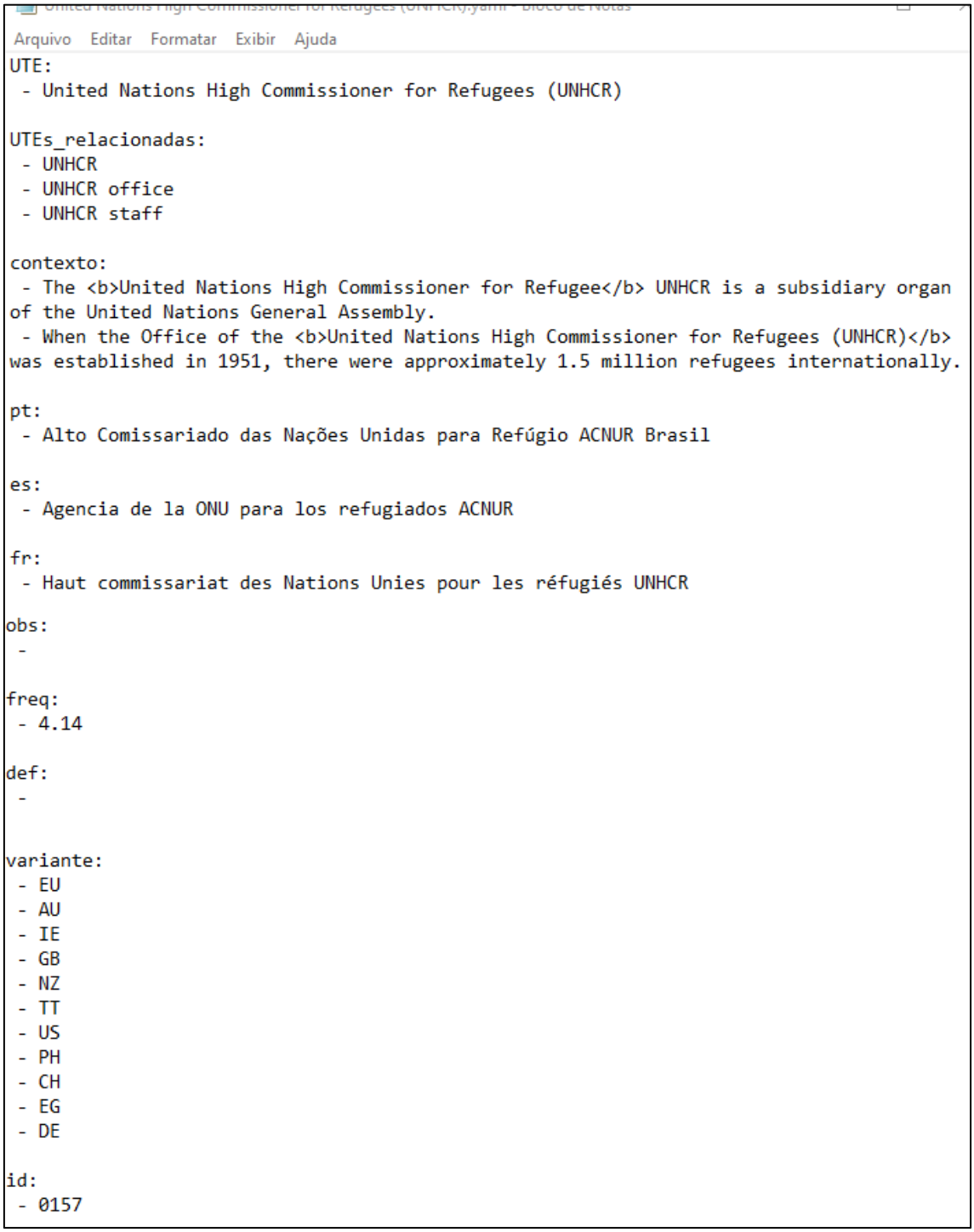
*UNHCR staff.yaml - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
UTE:
 - UNHCR staff

UTEs_relacionadas:
 - UNHCR
 - United Nations High Commissioner for Refugees (UNHCR)
 - UNHCR office

contexto:
 - Increasingly, <b>UNHCR staff</b> internally displaced persons (IDPs) and partners
are working in areas of armed conflict where there is little or no effective government
authority.
 - .<b>UNHCR staff</b> who receive oral requests should ensure that the information
required to support the request is received, and should record the details of the
request in writing on the individual file.
pt:
 - funcionário do ACNUR

es:
 - personal del ACNUR

fr:
 - não encontrada

obs:
 -

freq:
 - 3.40

def:
 -

variante:
 - EU
 - TT
 - CH
 - GB
 - EG

id:
 - 0145
```

Fonte: elaboração própria.

Figura 17 - Captura de tela da ficha de UTE "United Nations High Commissioner for Refugees"



Arquivo Editar Formatar Exibir Ajuda

```
UTE:
 - United Nations High Commissioner for Refugees (UNHCR)

UTEs_relacionadas:
 - UNHCR
 - UNHCR office
 - UNHCR staff

contexto:
 - The <b>United Nations High Commissioner for Refugee</b> UNHCR is a subsidiary organ
of the United Nations General Assembly.
 - When the Office of the <b>United Nations High Commissioner for Refugees (UNHCR)</b>
was established in 1951, there were approximately 1.5 million refugees internationally.

pt:
 - Alto Comissariado das Nações Unidas para Refúgio ACNUR Brasil

es:
 - Agencia de la ONU para los refugiados ACNUR

fr:
 - Haut commissariat des Nations Unies pour les réfugiés UNHCR

obs:
 -

freq:
 - 4.14

def:
 -


variante:
 - EU
 - AU
 - IE
 - GB
 - NZ
 - TT
 - US
 - PH
 - CH
 - EG
 - DE

id:
 - 0157
```

Fonte: elaboração própria.

A lista com todas as UTEs preenchidas neste trabalho, assim como o link para consulta online podem ser encontrados no Apêndice III. A seguir, mostramos as listas de

palavras-chave que usamos como ponto de partida neste trabalho e, em seguida, como funciona o sistema de busca de UTEs do Glossário Online.

7.3. Resultados extraídos do Sketch Engine

7.3.1 Listas de palavras-chave

Como dissemos na metodologia, utilizamos as listas de palavras-chave fornecidas pelo SK para criar as fichas-mãe. Tínhamos em mente explorar pelo menos as 10 primeiras palavras-chave dos quatro sub-corpus, mas a tarefa se tornou cada vez mais desafiadora, pois a combinação dos resultados obtidos pelos resumos do WordSketches com as informações do Concordance nos forneceu muito material para processar em tão pouco tempo. Por isso, segue abaixo a tabela de palavras-chaves que usamos neste trabalho.

Tabela 4 – Primeiras keywords de cada corpus, extraídas pelo Sketch Engine.

Português				Espanhol			
Keyword	**Score**	**Freq**	**Ref freq**	**Keyword**	**Score**	**Freq**	**Ref freq**
Acnur	2.151.590	3194	6	acnur	1.043.880	3080	10910
refugiados	1.510.460	2230	0	refugiado	935.370	10037	68618
estados	1.397.420	2063	0	reasentamiento	924.950	1865	3982
Conare	977.750	1443	0	refugiados	683.440	1968	10389
Direitos	969.630	1431	0	asilo	644.210	5160	48443
debates	839.660	1239	0	reasentar	592.300	962	1080
apátrida	718.520	1149	95	solicitante	289.640	438	269
cartagena	664.440	981	1	supra	194.780	287	0
américa	608.410	1006	137	asylum	171.030	256	181
solicitante	605.570	2310	1795	accem	166.080	3364	139323
Refúgio	559.650	3547	3731	apátrida	150.750	570	17132
refugiado	536.680	7623	9766	oua	142.810	253	2232
refugiar	512.970	3013	3375	refugiar	105.940	166	740
reassentamento	501.970	1045	466	immigration	104.180	247	6705
Latina	455.870	672	0	solicitantes	102.260	160	727
Brasília	444.010	684	51	law	101.350	1116	70794

México	420.610	638	33	conare	91.160	144	852

Francês				Inglês			
Keyword	**Score**	**Freq**	**Ref freq**	**Keyword**	**Score**	**Freq**	**Ref freq**
Asile	638.100	852	269	asylum	983.430	11303	102197
CGRA	485.900	6790	110988	UNHCR	966.630	4472	27571
DAR	426.070	576	424	Refugee	718.010	3061	23632
UNHCR	293.510	430	1427	Asylum	576.320	2341	21453
Ofpra	279.420	587	6998	RSD	508.900	1329	5696
ASILE	263.360	365	739	refugee	432.210	12328	287294
CNDA	246.150	328	274	Refugees	418.660	1379	13121
HCR	231.300	325	913	resettlement	329.450	1277	19464
OFPRA	226.960	927	24384	seeker	244.110	4313	169369
réfugier	194.140	313	2737	Migration	230.350	1101	29313
Exil	184.610	6341	289497	Immigration	188.290	1786	80473
subsidiaire	150.540	228	1898	stateless	173.910	527	10334
apatride	148.060	415	13185	unaccompanied	167.690	625	17911
Traite	127.280	290	8604	REFUGEES	158.020	332	247
demandeur	126.340	252	6118	Seekers	144.100	412	8500
SPR	118.060	1665	112192	Resettlement	142.100	330	2668
OE	116.020	2343	165560	OFPRA	138.660	288	6

Fonte: elaboração própria

É importante dizer que as listas foram obtidas diretamente do SK e não sofreram quaisquer alterações. Nas informações, **Keyword** corresponde a palavra-chave, **Score** é a chavicidade (por ordem decrescente), **Freq** equivale a frequência da palavra-chave no nosso corpus de estudo e **Ref freq** representa da frequência da palavra-chave no corpus de referência. No Apêndice III, mostramos uma tabela que lista quantas e quais fichas-mães e fichas-filhas foram produzidas neste trabalho. No Apêndice IV, apresentamos as listas de candidatos a termos extraídas com o SK, que são mais extensas.

8. Buscando UTEs no sistema de busca online

Para acessar o sistema, basta acessar o link http://bit.ly/glossariocommire e a página a seguir será exibida.

Figura 18 - Página inicial do glossário.

Fonte: elaboração própria.

Para fazer uma busca, basta digitar uma palavra ou expressão, em qualquer uma das línguas (português, inglês, francês e espanhol) no espaço reservado e clicar o botão "buscar", como indicado na Figura 18.

O sistema exibe os resultados para a expressão buscada, como exemplifica a Figura 18.

Figura 19 - Captura da tela de resultados por ordem alfabética de ocorrência no banco de dados

Fonte: elaboração própria.

8. CONSIDERAÇÕES FINAIS

Ao longo deste projeto, descrevemos o processo de compilação de um glossário multilíngue online sobre migrações e refúgio para tradutores e intérpretes usando a Linguística de Corpus como abordagem.

Para tanto, buscamos demonstrar o porquê de escolhermos utilizar a Linguística de Corpus como fundamentação teórica principal, e não a Terminologia ou a Lexicologia Especializada, como seria possível. Utilizamos a LC pois acreditamos ser a abordagem que possibilita uma maior representatividade das Unidade de Tradução Especializadas da área – refúgio e imigração – pelo fato de partir de textos autênticos aos quais imigrantes, refugiados, tradutores e intérpretes têm acesso cotidianamente.

Para empreender este projeto de um Glossário Multilíngue Online sobre Migração e Refúgio, utilizamos o COMMIRE, **Co**rpus **M**ultilíngue de **Mi**gração e **Re**fúgio composto de quatro sub-corpus em português, inglês, francês e espanhol, resultados de três anos de trabalho de iniciação científica junto aos grupos MOBILANG e TermTraDiCo.

Ressaltamos a necessidade de que este glossário fosse apresentado de forma online e gratuita, uma vez que não há no mercado uma opção de repertório lexicográfico ou terminográfico, impresso ou online, que seja multilíngue, gratuito e que tenha como público-alvo principal tradutores e intérpretes.

Para a exploração e extração dos dados do Glossário, utilizamos o programa Sketch Engine, que possibilitou a obtenção de uma grande quantidade de informações por meio das ferramentas WordSketches, Concordance e Keywords. Depois, registramos os dados obtidos em fichas de UTEs codificadas em utf-8 no formato YAML. Para criar o

banco de dados, utilizamos a linguagem de programação Python e o ambiente de desenvolvimento e hospedagem de aplicativos do Google, App Engine.

Em termos dos fatores limitantes para a realização dessas tarefas, podemos dizer que o tempo foi o maior empecilho, dado que tínhamos uma quantidade enorme de dados e pouco tempo para processá-los. Além disso, descobrimos que a extração de dados teria sido ainda mais proveitosa se o corpus fosse ainda maior. Outra dificuldade que tivemos foi no preenchimento das fichas, atividade que exigiu bastante atenção e demandou muito tempo – motivo pelo qual poucas foram produzidas até o momento.

No que concerne os próximos passos, sugerimos que todos os sub-corpus sejam limpos e transformados em arquivos de texto sem formatação, de forma a eliminar o ruído nos dados de forma consistente. Em seguida, será necessário que mais fichas sejam feitas, para que o glossário possa se apresentar de forma eficiente. Por fim, é preciso refinar as opções de busca, possibilitando a pesquisa em outros campos e a filtragem dos resultados, que esperamos poder apresentar em um site que seja multilíngue e acessível.

9. REFERÊNCIAS BIBLIOGRÁFICAS

ACNUR. **Refúgio em Números**. 3ª ed. 2018. Disponível em: https://www.acnur.org/portugues/wp-content/uploads/2018/04/refugio-em-numeros_1104.pdf. Acesso em: 03 jun 2019.

ANDRADE, Maria Margarida. **Lexicologia, Terminologia: definições, finalidades, conceitos operacionais**. In: Ana Maria Pinto de OLIVEIRA, Aparecida Negri ISQUERDO (Org.) As ciências do léxico: Lexicologia, Lexicografia, Terminologia. 2. ed. Campo Grande: Editora UFMS. 2001.

BARBOSA, 2001. **Dicionário, vocabulário, glossário: concepções**. In: ALVES, Ieda Maria (org.). A Constituição da normalização terminológica no Brasil. 2. ed. São Paulo: Humanitas FFLCH/USP. 2001.

BARROS, Lídia Almeida. **Curso Básico de Terminologia**. São Paulo: Editora Universidade de São Paulo, 2004.

BERBER SARDINHA, A. P. **Linguística de Corpus**. Barueri: Manole. 2004.

BOWKER, L. e PEARSON, J. **Working with Specialized Language**. London e New York: Routledge. 2002.

CASTELLS, Manuel. **A galáxia da Internet**: reflexões sobre a Internet, os negócios e a sociedade. Rio de Janeiro: Jorge Zahar, 2003.

CABRÉ, M. Teresa. **Terminology: Theory, methods and applications**. Ed. John Benjamins. 1999.

FURTADO, A. B. D. e GOROVITZ, S. ; **Primeiros Passos Para A Compilação De Um Corpus Terminológico Sobre Situações De Mobilidade: Coleta E Análise De Duas Cartilhas Informativas Multilíngues**. In: 23º Congresso de Iniciação Científica da Unb e 14º do DF, 2017, Brasília. Congresso de Iniciação Científica da Unb e Congresso de Iniciação Científica do DF, 2017.

GARCÍA, Fernanda de D. **O Papel do Intérprete Comunitário na Entrevista de Solicitação de Refúgio**. 2019. Dissertação (Mestrado em Estudos de Tradução) - Departamento de Línguas Estrangeiras e Tradução, Universidade de Brasília, Brasília, 2019.

GUERRA, Míriam M. e ANDRADE, Karylleila de S. **O léxico sob perspectiva: contribuições da Lexicologia para o ensino de línguas**. In: Domínios da Lingu@gem. v 6. n 1. 2012, p. 226 - 241.

HOFFMAN, Charlotte. The spread of English and the growth of multilingualism with English in Europe. In **English in Europe**: The Acquisition of a Third Language, Jasone Cenoz & Ulrike Jessner (eds), 1–21. Clevedon: Multilingual Matters. 2000.

JUBILUT, Liliana Lyra. **O Procedimento de Concessão de Refúgio no Brasil**. Ministério da Justiça. 18p. Agosto 2014.

KILGARRIFF, A., Pavel Rychlý, Pavel Smrž, David Tugwell. Itri-04-08 the sketch engine. Information Technology. 2004.

KILGARRIFF, Adam. Simple maths for keywords. In Proceedings of Corpus Linguistics Conference CL2009, Mahlberg, M., González-Díaz, V. & Smith, C. (eds.). UK: University of Liverpool. 2009.

KRIEGER, M. G. e FINATTO, M. J. B. **Introdução à Terminologia**: Teoria e Prática. São Paulo: Contexto. 2004.

LORENTE, M. **A Lexicologia como Ponto de Encontro entre a Gramática e a Semântica**. In: ISQUERDO, A. N. e KRIEGER, M. G. As ciências do léxico. vol. II. Campo Grande: UFMS, 2004. p. 19 – 30.

ORSI, Vivian. **Lexicologia: o que há por trás do estudo das palavras**? In: GONÇALVES, Adair Vieira e GÓIS, Marcos L. de Sousa (Org). Ciências da linguagem: o fazer científico. Vol. 1. Campinas, SP: Mercado de Letras, 2012.

MCENERY, Tony e HARDIE, Andrew. **Corpus Linguistics**: Method, Theory and Practice. USA: Cambridge University Press. 2012.

MILITÃO, Cinthia Duarte. **O processo de pedido de refúgio e a integração acadêmica de refugiados na Universidade de Brasília**. 2017. 50 f. Trabalho de Conclusão de Curso (Bacharelado em Línguas Estrangeiras Aplicadas) — Universidade de Brasília, Brasília, 2017.

MIRANDA, Jéssica Gonçalves de Araújo. **Presas estrangeiras no Brasil: barreiras linguísticas**. 2016. 52 f., il. Trabalho de Conclusão de Curso (Bacharelado em Línguas Estrangeiras Aplicadas) — Universidade de Brasília, Brasília, 2016.

MOLINA CABRERA, Marta Ingrith. **Migrações e impasses no acesso à saúde: traduzir-se é preciso**. 2017. 137 f., il. Dissertação (Mestrado em Estudos da Tradução) — Universidade de Brasília, Brasília, 2017.

SAUSSURE, F. **Curso de linguística geral**. 3 ed. São Paulo: Cultrix. 2006.

TAGNIN, Stella E. O. **A produção de glossários direcionados pelo corpus e orientados ao tradutor como metodologia de formação de tradutores**. In: Anais do X Encontro Nacional de Tradutores e IX Encontro Internacional de Tradutores. Ouro Preto. Set, 2009.

TAGNIN, Stella E. O. **Corpora na e para a Tradução**. In: Corpora na Tradução. VIANA, Vander e TAGNIN, Stella E. O. São Paulo: Hub Editorial. 2015.

TEIXEIRA, Elisa Duarte. **A Lingüística de Corpus a serviço do tradutor: proposta de um dicionário de culinária voltado para a produção textual**. 2008. Tese (Doutorado em Estudos Lingüísticos e Literários em Inglês) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2008. doi:10.11606/T.8.2008.tde-16022009-141747. Acesso em: 2018-05-22.

APÊNDICE I

LISTA DE CÓDIGOS DE PAÍSES

Código	País
AR	Argentina
AT	Áustria
AU	Austrália
BE	Bélgica
CA	Canadá
CH	Suíça
CL	Chile
CO	Colômbia
CY	Chipre
DE	Alemanha
DK	Dinamarca
EC	Equador
EG	Egito
ES	Espanha
EU	União Europeia
FR	França
FR	França

GB	Reino Unido
GT	Guatemala
IE	Irlanda
LA	América Latina
LU	Luxemburgo
MT	Malta
MX	México
NL	Holanda
NZ	Nova Zelândia
PA	Panamá
PE	Peru
PH	Filipinas
PT	Portugal
SE	Suécia
US	Estados Unidos
UY	Uruguai
ZA	África do Sul

APÊNDICE II

LISTA DE ARQUIVOS E SUAS RESPECTIVAS NACIONALIDADES

ARQUIVOS DO CORPUS EM ESPANHOL

Nome do Arquivo	Tokens	País de origem
ES_001	2.156	México
ES_002	1.976	Peru
ES_003	3.656	México
ES_004	2.066	Espanha
ES_005	3.158	Espanha
ES_006	1.675	Espanha
ES_007	16.039	Espanha
ES_008	8.479	Panamá
ES_009	5.227	Espanha
ES_010	857	Espanha
ES_011	122.630	União Europeia
ES_012	884	Colômbia
ES_013	1.070	Espanha
ES_014	3.767	União Europeia
ES_015	102.229	Equador
ES_016	919	Costa Rica
ES_017	1.021	Chile
ES_018	37.537	Espanha
ES_019	1.044	Argentina
ES_020	19.637	Espanha
ES_021	3.530	Equador
ES_023	1.339	Colômbia
ES_024	779	México
ES_025	190.655	México
ES_026	81.416	Espanha
ES_027	50.197	União Europeia
ES_028	2.105	Panamá
ES_029	34.320	União Europeia
ES_030	1.823	México
ES_031	1.713	Peru
ES_032	47.826	União Europeia
ES_033	3.961	América Latina
ES_034	59.633	América Latina

ES_035	984	Argentina
ES_036	14.857	Espanha
ES_037	52.541	União Europeia
ES_038	3.145	México
ES_039	15.065	Espanha
ES_040	2.739	Espanha
ES_041	21.269	Espanha
ES_042	8.676	Espanha
ES_043	1.670	Argentina
ES_044	1.498	América Latina
ES_045	1.439	América Latina
ES_046	546	México
ES_047	43.611	Espanha
ES_048	10.310	México
ES_050	35.025	América Latina
ES_051	1.277	México
ES_052	243	México
ES_053	884	México
ES_055	1.087	México
ES_057	3.724	Argentina
ES_060	1.822	Chile
ES_061	1.614	Chile
ES_062	5.460	Guatemala
ES_065	13.647	União Europeia
ES_067	416	Peru
ES_069	2.877	Peru
ES_070	7.752	México
ES_071	1.562	América Latina
ES_072	1.209	Uruguai
ES_074	768	Peru
ES_075	5.929	Estados Unidos
ES_077	7.565	Espanha
ES_078	2.219	Equador
ES_079	9.242	Espanha
ES_080	2.193	México
ES_081	5.822	Chile

ARQUIVOS DO CORPUS EM FRANCÊS

Nome do Arquivo	Tokens	País de origem
FR_001	29.363	Bélgica
FR_002	16.755	França
FR_003	15.866	França
FR_004	42.008	França
FR_005	14.197	França
FR_006	3.788	França
FR_007	2.439	Canadá
FR_008	991	Luxemburgo
FR_009	1.013	Canadá
FR_010	12.851	Canadá
FR_011	958	Canadá
FR_012	11.969	França
FR_013	1.910	Suíça
FR_014	558	Suíça
FR_015	1.058	Suíça
FR_016	2.963	Suíça
FR_017	2.800	Suíça
FR_018	10.180	Canadá
FR_019	3.207	Portugal
FR_020	9.663	Canadá
FR_021	3.069	Quebeque - Canadá
FR_022	3.826	Luxemburgo
FR_023	516	Canadá
FR_024	817	Canadá
FR_025	3.434	Luxemburgo
FR_026	332	Canadá
FR_027	35.544	Suíça
FR_028	3.971	Canadá
FR_029	33.071	Suíça
FR_030	3.717	Suíça
FR_031	1.802	Suíça
FR_032	6.213	Suíça
FR_033	8.276	Canadá
FR_034	88	Canadá
FR_035	1.125	Bélgica
FR_036	372	Canadá
FR_037	324	Canadá
FR_038	233	Canadá

FR_039	25.509	Canadá
FR_040	12.899	Luxemburgo
FR_041	74.473	Suíça
FR_042	1.267	Suíça
FR_043	14.540	França
FR_044	1.169	Canadá
FR_045	3.689	Canadá
FR_046	1.734	Canadá
FR_047	2.830	Canadá
FR_048	241	Canadá
FR_049	1.240	Canadá
FR_050	2.039	Suíça
FR_051	1.414	Bélgica
FR_052	143.173	Bélgica
FR_053	170	Canadá
FR_054	90	Canadá
FR_055	3.853	Bélgica
FR_056	5.035	Suíça
FR_057	4.752	Suíça
FR_058	4.544	Suíça
FR_059	4.508	Suíça
FR_060	4.476	Suíça
FR_061	4.757	Suíça
FR_062	4.375	Suíça
FR_063	4.320	Suíça
FR_064	4.512	Suíça
FR_065	4.717	Suíça
FR_066	4.500	Suíça
FR_067	4.361	Suíça
FR_068	4.673	Suíça
FR_069	4.754	Suíça
Fr_070	4.769	Suíça
FR_071	5.139	Suíça
FR_072	5.568	Suíça
FR_073	4.744	Suíça
FR_074	4.916	Suíça
FR_075	5.044	Suíça
FR_076	4.917	Suíça
FR_077	5336	Suíça

FR_078	5.115	Suíça
FR_079	4542	Suíça
FR_080	4.619	Suíça
FR_081	5.258	Suíça
FR_082	5.115	Suíça
FR_083	5.129	Suíça
FR_084	5.236	Suíça
FR_085	4.881	Suíça
FR_086	5.271	Suíça
FR_087	1.311	França
FR_088	1.253	Chipre
FR_089	4.789	Holanda
FR_090	1.837	França
FR_091	2.292	França
FR_092	954	França
FR_093	1.396	Áustria
FR_094	2.166	França
FR_095	3.660	União Europeia
FR_096	9.776	França
FR_097	2.980	Suíça
FR_098	1.347	Suíça
FR_099	4.839	União Europeia
FR_100	18.146	França
FR_101	30.126	França
FR_102	3.609	União Europeia
FR_103	1.347	Alemanha
FR_104	1.247	Canadá
FR_105	14.504	União Europeia
FR_106	1.524	França
FR_107	1.004	Dinamarca
FR_108	29.768	França
FR_109	1.307	União Europeia
FR_110	3.203	Bélgica
FR_111	4.262	França
FR_112	4.400	Bélgica
FR_113	4.099	Bélgica
FR_114	32.887	Suíça
FR_115	82.729	França
FR_116	11.275	França

FR_117	13.873	França
FR_118	6.137	França
FR_119	3.221	França
FR_120	12.959	França
FR_121	9.270	França
FR_122	4.844	França
FR_123	5.341	Canadá
FR_124	10.834	Suíça
FR_125	4.689	Bélgica
FR_126	6.484	Bélgica
FR_127	3.646	Bélgica
FR_128	3.435	Bélgica
FR_129	4.961	Bélgica
FR_130	3.782	França
FR_131	4.892	França

ARQUIVOS DO CORPUS EM INGLÊS

Nome do Arquivo	Tokens	País de origem
EN_001	24.508	Estados Unidos
EN_002	6.417	Estados Unidos
EN_003	39.157	União Europeia
EN_004	9.234	Canadá
EN_005	2.538	Canadá
EN_006	10.273	Canadá
EN_007	110.104	União Europeia
EN_008	19.587	Canadá
EN_009	3.064	Canadá
EN_010	1.284	Canadá
EN_011	15.741	Austrália
EN_012	1.452	Austrália
EN_013	32.981	Estados Unidos
EN_014	6.271	Austrália
EN_015	13.555	Austrália
EN_016	4.763	Austrália
EN_017	20.866	Austrália
EN_018	14.664	União Europeia
EN_019	4.375	Reino Unido
EN_020	103.493	Irlanda

EN_021	9.623	Irlanda
EN_022	6.018	Irlanda
EN_023	16.257	União Europeia
EN_024	10.747	União Europeia
EN_025	48.531	Irlanda
EN_026	14.913	Escócia (Reino Unido)
EN_027	8.676	Reino Unido
EN_028	6.806	Escócia (Reino Unido)
EN_029	44.308	União Europeia
EN_030	4.298	Reino Unido
EN_031	16.301	Nova Zelândia
EN_032	1.797	Nova Zelândia
EN_033	7.669	Nova Zelândia
EN_034	39.068	Nova Zelândia
EN_035	17.280	Nova Zelândia
EN_036	30.914	Nova Zelândia
EN_037	231.739	União Europeia
EN_038	671	Nova Zelândia
EN_039	5.087	Nova Zelândia
EN_040	2.852	Nova Zelândia
EN_041	3.025	Nova Zelândia
EN_042	1.765	Nova Zelândia
EN_043	1.302	Nova Zelândia
EN_044	1.376	Nova Zelândia
EN_045	1.249	Nova Zelândia
EN_046	1.040	Nova Zelândia
EN_047	5.774	Nova Zelândia
EN_048	30.219	Nova Zelândia
EN_049	3.073	União Europeia
EN_050	8.129	União Europeia
EN_051	45.791	Austrália
EN_052	19.249	Suécia
EN_053	10.951	União Europeia
EN_054	17.822	União Europeia
EN_055	2.543	África do Sul
EN_056	4.572	África do Sul
EN_057	4.742	África do Sul
EN_058	47.737	África do Sul
EN_059	806	Trindade e Tobago

EN_060	28.703	Trindade e Tobago
EN_061	30.526	Estados Unidos
EN_062	21.544	Alemanha
EN_063	949	Alemanha
EN_064	2.571	Alemanha
EN_065	21.544	Alemanha
EN_066	5.179	Filipinas
EN_067	2.194	Filipinas
EN_068	52.610	Reino Unido
EN_069	43.289	Suíça
EN_070	1.261	Malta
EN_071	29.630	Malta
EN_072	5.703	Inglaterra (Reino Unido)
EN_073	4.559	Suíça
EN_074	6.363	Suíça
EN_075	4.719	Suíça
EN_076	3.995	Suíça
EN_077	11.297	Suíça
EN_078	17.165	França
EN_079	15.331	França
EN_080	5.363	Egito
EN_081	6.366	Reino Unido
EN_082	21.158	Egito
EN_083	17.897	Reino Unido
EN_084	14.330	Reino Unido
EN_085	210	Canadá
EN_086	293	Canadá
EN_087	189	Canadá
EN_088	139	Canadá
EN_089	66	Canadá
EN_090	297	Canadá
EN_091	67	Canadá
EN_092	277	Canadá
EN_093	4.000	Reino Unido
EN_094	9.313	Reino Unido
EN_095	53.523	Inglaterra (Reino Unido)
EN_096	25.295	Reino Unido
EN_097	8.884	Alemanha
EN_098	4.721	Inglaterra (Reino Unido)

EN_099	51.470	Estados Unidos
EN_100	5.136	Estados Unidos
EN_101	52.807	Estados Unidos

APÊNDICE III

LISTA DE UTES PREENCHIDAS NESTE TRABALHO

Num.	UTE/p. de busca	Língua	Tipo
1	**asylum**	EN	p.chave
2	asylum application	EN	UTE
3	asylum procedure	EN	UTE
4	asylum seeker	EN	UTE
5	asylum claim	EN	UTE
6	seek asylum	EN	UTE
7	asylum process	EN	UTE
8	asylum system	EN	UTE
9	asylum policy	EN	UTE
10	asylum applicant	EN	UTE
11	asylum interview	EN	UTE
12	asylum support	EN	UTE
13	asylum officer	EN	UTE
14	asylum case	EN	UTE
15	asylum law	EN	UTE
16	asylum status	EN	UTE
17	asylum country	EN	UTE
18	asylum determination process	EN	UTE
19	grant asylum	EN	UTE
20	destitute asylum seekers	EN	UTE
21	asylum request	EN	UTE
22	failed asylum seeker	EN	UTE
23	vulnerable asylum seeker	EN	UTE
24	examine na asylum application	EN	UTE
25	submit an asylum application	EN	UTE
26	process an asylum application	EN	UTE
27	regular asylum procedure	EN	UTE
28	individual asylum procedure	EN	UTE
29	process an asylum claim	EN	UTE
30	make an asylum claim	EN	UTE
31	refuse an asylum claim	EN	UTE
32	lodge an asylum claim	EN	UTE
33	reject an asylum claim	EN	UTE
34	submit an asylum claim	EN	UTE
35	national asylum system	EN	UTE
36	country of first asylum	EN	UTE
37	asylum claimant	EN	UTE
38	asylum authorities	EN	UTE

39	asylum proceedings	EN	UTE
40	asylum seeker children	EN	UTE
41	asylum flows	EN	UTE
42	claim asylum	EN	UTE
43	request asylum	EN	UTE
44	get asylum	EN	UTE
45	apply for asylum	EN	UTE
46	asylum office	EN	UTE
47	**UNHCR**	EN	p.chave
48	United Nations High Commissioner for Refugees (UNHCR)	EN	UTE
49	UNHCR office	EN	UTE
50	UNHCR staff	EN	UTE
51	determinación de la condición de refugiado	ES	UTE
52	solicitantes de la condición de refugiado	ES	UTE
53	**refugiado**	ES	p.chave
54	menor refugiado	ES	UTE
55	solicitud de reconocimiento de la condición de refugiado	ES	UTE
56	**asilo**	ES	p.chave
57	asilo diplomático	ES	UTE
58	asilo territorial	ES	UTE
59	asilo político	ES	UTE
60	solicitante de asilo político	ES	UTE
61	país de primer asilo	ES	UTE
62	país de asilo	ES	UTE
63	solicitar asilo	ES	UTE
64	buscar asilo	ES	UTE
65	solicitud de asilo	ES	UTE
66	solicitante de asilo	ES	UTE
67	proceso de asilo	ES	UTE
68	sistema de asilo	ES	UTE
69	política de asilo	ES	UTE
70	oficina de asilo	ES	UTE
71	ley de asilo	ES	UTE
72	conceder asilo	ES	
73	procedimiento de reconocimiento de la condición de refugiado	ES	UTE
74	demandante de asilo	ES	UTE
75	obtener asilo	ES	UTE
76	presentar una solicitud de asilo	ES	UTE
77	derecho de buscar y recibir asilo	ES	UTE
78	recibir asilo	ES	UTE
79	pedir asilo	ES	UTE
80	otorgar asilo	ES	UTE
81	denegar asilo	ES	UTE
82	derecho de asilo	ES	UTE

83	procedimiento de asilo	ES	UTE
84	en materia de asilo	ES	UTE
85	persona solicitante de asilo	ES	UTE
86	otorgamiento de asilo	ES	UTE
87	institución del asilo	ES	UTE
88	**refugiado**	ES	p.chave
89	asilo	ES	UTE
91	derecho a buscar y obtener asilo	ES	UTE
92	status de refugiado	ES	UTE
93	**ACNUR**	ES	p.chave
94	Agencia de la ONU para los refugiados ACNUR	ES	UTE
95	oficina del ACNUR	ES	UTE
96	asile	FR	UTE
97	demander asile	FR	UTE
98	requérant d'asile débouté	FR	UTE
99	chercher asile	FR	UTE
100	obtenir l'asile	FR	UTE
101	accorder l'asile	FR	UTE
102	déposer une demande d'asile	FR	UTE
103	présenter une demande d'asile	FR	UTE
104	requérants d'asile mineurs non accompagnés	FR	UTE
105	système d'asile national	FR	UTE
106	demandeur d'asile	FR	UTE
107	procédure d'asile	FR	UTE
108	pays d'asile	FR	UTE
109	demande d'asile	FR	UTE
110	système d'asile	FR	UTE
111	politique d'asile	FR	UTE
112	statut de refugié	FR	UTE
113	examiner une demande d'asile	FR	UTE
114	faire une demande d'asile	FR	UTE
115	officier de protection	FR	UTE
116	traiter une demande d'asile	FR	UTE
117	procédure normale	FR	UTE
118	pays de premier asile	FR	UTE
119	formulaire de demande d'asile	FR	UTE
120	droit d'asile	FR	UTE
121	**HCR**	FR	p.chave
122	Haut-commissariat des Nations Unies pour les réfugiés UNHCR	FR	UTE
123	status de refugiado	PT	UTE
124	**refúgio**	PT	p.chave
125	solicitante de refúgio	PT	UTE
126	solicitação de refúgio	PT	UTE
127	pedido de refúgio	PT	UTE
128	lei de refúgio	PT	UTE

129	país de refúgio	PT	UTE
130	concessão de refúgio	PT	UTE
131	instituto do refúgio	PT	UTE
132	procedimento de refúgio	PT	UTE
133	processo de refúgio	PT	UTE
134	situação de refúgio	PT	UTE
135	pessoas em situação de refúgio	PT	UTE
136	direito de refúgio	PT	UTE
137	política de refúgio	PT	UTE
138	reconhecimento de refúgio	PT	UTE
139	direito de buscar e receber refúgio	PT	UTE
140	reconhecimento da condição de refúgio	PT	UTE
141	em busca de refúgio	PT	UTE
142	proteção do refúgio	PT	UTE
143	solicitar refúgio	PT	UTE
144	buscar refúgio	PT	UTE
145	receber refúgio	PT	UTE
146	conceder refúgio	PT	UTE
147	pedir refúgio	PT	UTE
148	obter refúgio	PT	UTE
149	circunstâncias que determinaram o refúgio	PT	UTE
150	sistema de refúgio	PT	UTE
151	primeiro país de refúgio	PT	UTE
152	formulário de solicitação de refúgio	PT	UTE
153	termo de solicitação de refúgio	PT	UTE
154	protocolo de solicitação de refúgio	PT	UTE
155	apresentar uma solicitação de refúgio	PT	UTE
156	solicitante de reconocimiento de la condición de refugiado	ES	UTE
157	**ACNUR**	PT	p.chave
159	funcionário do ACNUR	PT	UTE
160	escritório do ACNUR	PT	UTE
161	Alto Comissariado das Nações Unidas para Refúgio ACNUR Brasil	PT	UTE

As fichas desta lista encontram-se disponíveis em: [http://bit.ly/commire-fichas]

APÊNDICE IV

LISTAS DE CANDIDATOS A TERMO FORNECIDA PELO SK

CEM PRIMEIROS CANDIDATOS DO CORPUS DO INGLÊS

Term	Score	Freq	Ref freq
refugee status	538.500	1408	60
asylum application	321.900	688	6
asylum procedure	278.950	586	2
international protection	235.440	511	10
particular social group	201.600	444	14
asylum seeker	190.910	503	63
residence permit	186.450	460	44
subsidiary protection	175.770	367	1
social group	172.630	520	106
refugee protection	149.430	322	9
asylum claim	148.250	321	10
third country	147.910	371	49
status determination	108.750	229	4
political opinion	106.620	249	30
family reunification	105.500	251	35
seeking asylum	101.890	237	29
immigration detention	94.100	226	38
well-founded fear	85.520	185	11
asylum process	82.100	171	2
travel document	78.470	175	19
safe third country	77.360	161	2
reception centre	72.990	153	4
temporary protection	70.540	154	14
international refugee	69.470	148	8
deportation order	69.260	150	12
negative decision	68.280	143	4
immigration officer	67.850	154	24
sexual health	67.540	356	365
refugee status determination	66.660	139	3
principal applicant	63.830	133	3
safe country	59.240	129	14
south afric	58.850	121	0
t iz	58.850	121	0

protection status	57.460	120	4
refugee law	56.510	122	12
1s t	55.980	115	0
suspensive effect	54.550	112	0
host country	53.920	169	123
family reunion	53.730	193	175
protected person	52.960	111	5
first instance	52.840	253	311
legal assistance	52.760	176	146
permanent residence	52.220	153	100
asylum system	51.890	115	19
complementary protection	51.680	106	0
refugee claim	50.320	104	2
refugee definition	49.770	102	0
immigration court	48.180	102	8
habitual residence	47.900	101	7
regular procedure	47.810	100	5
judicial review	47.620	198	240
degrading treatment	47.040	114	43
legal aid	46.560	203	263
origin information	46.050	95	2
stateless person	46.050	95	2
country of origin information	44.160	91	2
voluntary return	44.090	94	10
fpa p	43.080	88	0
national legislation	42.390	116	79
european convention	41.540	106	58
home country	40.840	258	486
derivative status	40.670	84	3
work permit	40.470	106	66
asylum policy	38.770	82	9
removal order	38.310	79	3
asylum office	36.860	75	0
legal representative	36.810	93	56
immigration status	36.670	123	150
normal procedure	36.470	79	15
public order	36.220	102	89
health screening	36.050	85	37
regional processing	35.140	72	2

local integration	35.000	72	3
labour market	35.000	173	331
rsd decision	34.950	71	0
need of international protection	34.950	71	0
humanitarian status	34.470	70	0
voluntary repatriation	34.230	71	5
serious harm	33.730	88	66
asylum applicant	33.510	68	0
asylum interview	33.510	68	0
irregular migration	32.900	69	8
family unity	32.830	72	19
international refugee law	31.950	65	1
asylum support	31.120	63	0
asylum officer	29.690	60	0
mixed migration	28.730	58	0
mass influx	27.920	57	3
genital mutilation	27.710	78	91
refugee family	26.500	54	3
personal interview	26.370	63	43
temporary residence	26.130	56	15
such person	26.010	91	169
armed conflict	25.720	111	262
subsequent application	25.700	53	6
many asylum	25.660	52	2
other humanitarian status	25.380	51	0
social assistance	25.330	62	50
durable solution	25.040	52	8

CEM PRIMEIROS CANDIDATOS DO CORPUS DO ESPANHOL

Term	Score	Freq	Ref freq
condición de refugiado	891.410	1162	30
protección internacional	875.320	1141	80
país de origen	710.570	926	1453
solicitante de asilo	637.300	1020	56
derecho del niño	230.880	300	557
convención americana	204.830	266	228
solicitud de asilo	200.440	288	26
necesidad de protección	197.930	257	79

reconocimiento de la condición	197.170	256	10
protección complementaria	194.870	253	7
país de asilo	192.570	250	11
unión europea	189.500	246	9971
estatuto de el refugiado	186.500	254	12
país de reasentamiento	181.070	235	11
relación exterior	180.310	234	1407
opinión consultiva	175.710	228	41
comisionado de la nación	173.410	225	96
sociedad receptora	167.800	256	43
derecho de asilo	165.980	240	28
estatuto de refugiado	157.280	214	12
declaración de cartagena	156.550	203	12
protección de el refugiado	155.790	202	13
situación de refugio	153.490	199	8
comité ejecutivo	141.230	183	1335
programa de reasentamiento	140.360	187	7
opinión política	139.700	181	164
determinación de la condición	137.400	178	19
interés superior	129.730	168	205
solución duradera	123.600	160	38
necesidad de protección internacional	119.770	155	2
definición de refugiado	119.770	155	2
consejo de la judicatura federal	119.470	173	29
judicatura federal	115.900	174	39
w w	114.620	158	16
país receptor	112.880	146	134
américa latina	110.580	143	9489
medio de comunicación	108.280	140	13794
formación lingüística	107.510	139	37
orden público	107.510	139	1001
dirección general	106.750	138	6268
comité de el derecho	105.210	136	31
observación general	103.680	134	40
condición jurídica	100.620	130	78
protección subsidiaria	100.620	130	2
convención de ginebra	99.850	129	49
programa de integración	99.850	129	117
comunidad de refugiado	99.790	131	4

reasentamiento de refugiado	97.930	127	1
procedimiento de asilo	97.540	127	2
oficina de asilo	94.510	124	4
ministerio de relación	94.480	122	546
ministerio de relación exterior	94.480	122	501
día hábil	92.190	119	2072
declaración americana	91.420	118	45
reunificación familiar	90.650	117	41
convención de la oua	89.890	116	0
derecho de el refugiado	89.680	119	7
asilo político	86.820	112	79
país de acogida	86.060	111	156
legislación nacional	86.060	111	380
consejo de la judicatura	85.890	173	136
organización internacional	85.290	110	1506
ciudad de méxico	84.520	109	3824
solicitud de protección	83.760	108	8
representación del acnur	82.990	107	0
ley de asilo	82.660	107	1
estado contratante	81.090	117	29
persona colombiana	78.840	102	1
país de su nacionalidad	78.390	101	5
persona migrante	75.640	128	76
situación migratoria	75.330	97	46
trabajador migrante	75.050	141	111
comité ejecutivo del acnur	74.560	96	0
declaración universal	73.800	95	428
proceso de reasentamiento	73.560	97	6
comisión interamericana	72.260	93	336
república dominicana	72.260	93	2166
situación de conflicto	71.500	92	226
protección de el derecho	71.500	92	621
orientación sexual	70.730	91	729
solicitud de protección internacional	70.730	91	1
instrumento internacional	70.020	183	249
misión de selección	69.960	90	3
solicitante de refugio	69.750	93	9
solicitante de protección	67.670	87	0
observación general número	66.130	85	2

niño refugiado	65.370	84	18
definición regional	64.600	83	0
guía práctica	64.600	83	324
persona solicitante	63.540	87	16
derecho internacional	63.130	541	1367
residencia habitual	63.100	120	116
situación irregular	63.070	81	224
motivo de género	60.770	78	12
w w w	60.570	79	4
estado contratante	60.560	88	32
protección temporal	60.000	77	12
enlace del acnur	59.240	76	0
unión interparlamentaria	59.240	76	24

CEM PRIMEIROS CANDIDATOS DO CORPUS DO FRANCÊS

Term	Score	Freq	Ref freq
protection internationale	555.560	877	50
êtres humains	362.250	472	2133
traite des êtres	308.680	402	68
traite des êtres humains	306.380	399	68
protection subsidiaire	266.610	385	26
victimes de la traite	230.940	346	36
union européenne	229.080	298	6829
mise en œuvre	201.530	262	4161
titre de séjour	189.280	246	211
organisation suisse	164.130	214	1
victimes de traite	152.540	198	0
mineurs isolés	148.720	193	3
demande de protection	130.280	181	17
isolés étrangers	124.990	162	0
demande de protection internationale	118.870	154	0
admission provisoire	118.680	155	2
mineurs isolés étrangers	118.100	153	0
social n	116.110	151	1
cahiers du social n	115.040	149	0
comité exécutif	113.510	147	501
groupe social	111.670	225	131
pays tiers	100.990	294	294
protection des victimes	97.860	141	27
statut de protection	96.170	129	9
officier de protection	95.140	124	2
intérêt supérieur	91.310	118	192

société civile	89.020	115	2925
carte de séjour	86.720	112	190
solutions durables	86.670	151	82
cas échéant	83.660	108	1790
état civil	79.830	103	748
contexte de migration	79.510	103	1
détermination du statut	79.070	102	23
assemblée générale	76.010	98	3197
besoins spécifiques	73.710	95	357
migration de transit	73.710	95	0
marché du travail	72.940	94	1270
réunification familiale	72.180	93	5
principe de non-refoulement	72.100	96	8
office fédéral	69.120	89	173
traite des personnes	68.820	103	38
reconnaissance du statut	64.600	88	14
lieu de résidence	64.530	83	337
regroupement familial	63.470	156	213
résidence habituelle	62.990	81	56
besoin de protection	60.690	94	48
séjour temporaire	59.930	77	48
cas de retour	59.340	87	33
autorisation de travail	59.300	84	24
résidence permanente	58.400	75	72
statut de protection subsidiaire	57.410	74	1
ressortissants de pays tiers	57.370	78	14
ressortissants de pays	57.080	82	28
personnes en quête	56.110	72	30
quête de protection	55.340	71	5
dépôt de la demande	54.580	70	51
entretien personnel	54.520	72	7
attestation de demande	53.050	68	0
permis de séjour	53.050	68	69
territoire français	51.510	66	577
opinions politiques	51.490	107	145
autorisation de séjour	49.980	64	17
législation nationale	49.980	64	169
accueil des demandeurs	49.980	70	22
principes directeurs	49.780	93	107
solution durable	49.660	92	104
autorités compétentes	49.220	63	453
le-la jeune	49.220	63	0
personnes vulnérables	49.220	63	85
même temps	49.220	63	18053
victime de traite	49.220	63	0

décision négative	47.690	61	10
exploitation sexuelle	46.990	96	139
besoins de protection	46.700	63	13
personne de confiance	46.400	82	89
code pénal	46.160	59	518
procédure accélérée	46.160	59	22
titre de voyage	45.430	59	4
carte de résident	44.630	57	50
liste de contrôle	44.630	57	133
consultation juridique	44.280	58	6
autorités nationales	43.860	56	149
mariage forcé	43.860	56	73
pays européen	43.860	56	234
soins médicaux	43.090	55	534
pôle emploi	42.330	54	556
situation irrégulière	42.330	54	274
traitements inhumains	42.220	65	48
représentant légal	41.560	53	96
représentation juridique	41.560	53	23
secrétaire général	41.560	53	3745
sécurité sociale	41.560	53	2512
office français	40.800	52	29
titres de voyage	40.620	54	10
titre provisoire	40.330	74	102
dépôt de plainte	39.270	50	69
parlement européen	39.270	50	1659
procédure normale	39.270	50	38
garanties procédurales	38.500	49	10

CEM PRIMEIROS CANDIDATOS DO CORPUS DO PORTUGUÊS

Term	Score	Freq	Ref freq
condição de refugiado	968.270	1429	20
caderno de debates	818.340	1222	3
proteção internacional	755.730	1115	221
país de origem	595.980	879	725
américa latina	455.190	671	13049
antónio guterres	390.210	575	6
resolução normativa	290.030	427	113
declaração de cartagena	285.970	421	20
plano de ação	282.580	416	1176
lei nº	262.280	386	0
tráfico de pessoas	261.600	385	200

marta juárez	260.250	383	0
polícia federal	252.800	372	6553
solicitantes de refúgio	244.000	359	8
integração local	234.530	345	9
temor de perseguição	234.530	345	7
migrações internacionais	204.740	301	35
solicitação de refúgio	197.970	291	3
orientação sexual	193.910	285	1119
proteção nacional	193.910	285	11
residência habitual	187.140	275	117
território nacional	185.790	273	3082
cançado trindade	181.050	266	70
instrumentos de proteção	179.020	263	39
políticas públicas	176.990	260	8373
soluções duradouras	175.640	258	15
resolução normativa nº	174.280	256	0
instrumentos de proteção nacional	172.250	253	0
coletânea de instrumentos	170.900	251	0
san josé	167.510	246	210
comitê nacional	157.360	231	252
brasil contemporâneo	149.240	219	149
paulo abrão	147.210	216	18
identidade de gênero	147.210	216	202
reassentamento solidário	143.150	210	1
vítimas de tráfico	138.410	203	31
países andinos	137.050	201	36
relações exteriores	136.380	200	2437
josé eduardo	133.670	196	1135
ponte neto	132.320	194	0
josé eduardo cardozo	131.640	193	361
antônio cançado	130.280	191	1
antônio cançado trindade	130.280	191	0
repatriação voluntária	128.930	189	20
sally valladares	128.250	188	0
jan egeland	127.570	190	4
carlos maldonado	127.070	190	5
rocha paranhos	126.560	187	2
consultas sub-regionais	126.220	185	0
feliciano ponte neto	125.550	184	0

feliciano ponte	125.550	184	0
eleonora menicucci	125.080	187	5
ato inaugural agenda	124.870	183	0
mercosul agenda	124.870	183	0
mesoamérica agenda	124.870	183	0
países andinos agenda	124.870	183	0
agenda discursos alberto	124.870	183	0
discursos alberto	124.870	183	0
torella antónio	124.870	183	0
cartagena apresentação	124.870	183	0
menicucci josé eduardo cardozo	124.870	183	0
abrão declaração	124.870	183	0
maldonado viii anexos declaração	124.870	183	0
paulo abrão declaração	124.870	183	0
viii anexos declaração	124.870	183	0
informação discurso	124.870	183	0
agenda discursos	124.870	183	0
ato inaugural agenda discursos	124.870	183	0
encerramento discursos	124.870	183	0
eleonora menicucci josé eduardo	124.870	183	0
menicucci josé eduardo	124.870	183	0
antónio guterres jan egeland	124.870	183	0
guterres jan egeland	124.870	183	0
vii participantes epílogo	124.870	183	0
iii evento	124.870	183	0
iii evento comemorativo	124.870	183	0
agenda discursos alberto figueiredo	124.870	183	0
discursos alberto figueiredo	124.870	183	0
torella antónio guterres	124.870	183	0
juárez introdução	124.870	183	0
marta juárez introdução	124.870	183	0
antónio guterres jan	124.870	183	0
guterres jan	124.870	183	0
eleonora menicucci josé	124.870	183	0
menicucci josé	124.870	183	0
discursos marta juárez	124.870	183	0
encerramento discursos marta juárez	124.870	183	0
discursos alberto figueiredo machado	124.870	183	0
discursos marta	124.870	183	0

encerramento discursos marta	124.870	183	0
iv rumo	124.870	183	0
vii participantes	124.870	183	0
carlos maldonado viii	124.870	183	0
carlos maldonado viii anexos	124.870	183	0
maldonado viii	124.870	183	0
maldonado viii anexos	124.870	183	0
viii anexos	124.870	183	0
grupo social específico	124.190	182	6
alberto figueiredo machado	123.200	188	10